Integrating Explainable AI in Medical Text Classification with Supervised Models

Bachelor Semester Project S4 (Academic Year 2024/25), University of Luxembourg

Ayoub MERDAN
FSTM
University of Luxembourg
Student

Salima LAMSIYAH
FSTM
University of Luxembourg
Tutor

ABSTRACT

Medical text classification is a paramount task in clinical natural language processing (NLP), allowing automated structuring of unstructured medical records, abstracts, and clinical notes. While machine learning and deep learning methods have achieved strong predictive performance, their widespread adoption in healthcare remains limited by the *black-box* nature of most models. In this project, we investigate the integration of **Explainable AI (XAI)** methods to enhance the interpretability of supervised classifiers for medical text. We evaluate traditional classifiers—Naïve Bayes, Decision Trees, Logistic Regression, and Support Vector Machines—in combination with text embedding approaches including TF–IDF, Word2Vec, and SBERT.

To bridge predictive performance with interpretability, we apply both *global* and *local* explanation techniques: top-likelihood token analysis for Naïve Bayes, coefficient visualization for linear models, tree path inspection for Decision Trees, and LIME/SHAP-based attribution for embedding-based models. Our experiments on a biomedical abstracts dataset highlight that domain-specific embeddings (e.g., SBERT) substantially improve classification accuracy, while interpretable models and post-hoc explanations provide valuable clinical insight into decision-making.

The results demonstrate that combining robust text representations with transparent interpretability techniques allows models to not only predict medical conditions effectively but also justify their outputs in ways that are clinically meaningful. This contributes toward building trustworthy AI systems for healthcare, where reliability and transparency are as critical as predictive power.

1 INTRODUCTION

Large volumes of clinical text—such as medical abstracts, discharge summaries, and progress notes—contain key information for downstream care and research. Automatic *medical text classification* can surface relevant cases, route documents, and support evidence synthesis. Despite strong accuracy from modern NLP models, deployment in healthcare is often hindered by the *black-box* nature of these systems: clinicians and auditors need to understand *why* a model produced a given label, not only *what* it predicted. Explainable AI (XAI) is therefore a prerequisite for reliability, error analysis, and regulatory readiness.

This work studies explainability for supervised classifiers trained on different text representations. We consider four widely-used classifiers—Naïve Bayes (NB), Decision Trees (DT), Logistic Regression (LR), and linear Support Vector Machines (SVM)—paired with both sparse and dense embeddings, including TF–IDF and Word2Vec, and contextual encoders such as SBERT (and, where applicable, domain-specific variants). These choices reflect common points on the performance–interpretability spectrum: linear models and shallow trees are inherently transparent on TF–IDF features, whereas dense and contextual embeddings typically require post-hoc explanations.

Goals and questions. We aim to connect predictive performance with human-understandable rationales. Concretely, we ask:

- **RQ1:** How do explanations differ across model families and text representations?
- **RQ2:** Can we obtain *global* summaries (e.g., important terms per class) for transparent models, and *local* token-level rationales for opaque pipelines?
- **RQ3:** Are the explanations *faithful* in the sense that removing highly-attributed tokens reduces the model's confidence?

Contributions. This paper makes the following contributions:

- **Unified XAI pipeline.** We implement a consistent framework that selects explanation methods by model type: NB term likelihoods and class-specific top words; LR/SVM coefficient visualizations; shallow tree plots and decision paths for DT; and model-agnostic, token-level attributions (LIME/SHAP) for Word2Vec/SBERT pipelines.
- **Qualitative and quantitative analysis.** We pair global views (per-class term summaries, coefficient bars) with local explanations on representative samples, and include lightweight faithfulness checks (deletion tests) to avoid purely anecdotal narratives.
- **Practical guidance.** From the comparative study, we distill recommendations for choosing classifiers and explainers under common constraints (accuracy needs, compute budget, and interpretability requirements in clinical settings).

Summary of findings. In our experiments on biomedical abstracts, contextual representations (e.g., SBERT) tend to improve accuracy over traditional TF–IDF and Word2Vec, while transparent models on TF–IDF yield the clearest *global* explanations. Post-hoc methods (LIME/SHAP) recover *local* rationales for dense/transformer embeddings, enabling token-level insight even when parameters are not directly interpretable.

2 RELATED WORK

2.1 Explainability for clinical NLP.

As machine learning models for clinical text have grown in complexity, a parallel literature has focused on explaining predictions to promote safety, auditability, and human oversight. Model-agnostic local explanation tools such as LIME [1] and SHAP [2] have become popular because they apply uniformly to linear models (e.g., Logistic Regression, linear SVM), probabilistic models (e.g., Naïve Bayes), trees, and deep networks, returning token-level contributions for individual decisions. In the healthcare NLP space specifically, recent surveys and scoping reviews document a rapid rise of XAI techniques for tasks ranging from document classification to diagnosis support, while stressing clinical validation and careful communication of uncertainty [3, 4].

2.2 Local vs. global explanations.

Local post-hoc methods (LIME, SHAP) explain *one* prediction at a time; global views can be built by aggregating local attributions or by inspecting model parameters directly (e.g., coefficient bars for LR/SVM, likelihood terms for Naïve Bayes, or feature importances and paths for trees). For deep encoders and transformers, gradient-based attribution such as Integrated Gradients (IG) provides token-level saliency with axiomatic guarantees of sensitivity and implementation invariance [5]. Clinical pipelines combining BERT-family encoders with attributions have been proposed for diagnosis from dialogues and reports, with radiologists assessing whether highlighted tokens align with domain reasoning [6, 7].

2.3 Faithfulness versus plausibility.

A central theme in XAI is that visually convincing highlights need not reflect the model's true internal evidence. For NLP, a line of work shows that attention weights are often *not* faithful explanations of model reasoning [8, 9]. Systematic evaluations in clinical settings likewise emphasize measuring the *faithfulness* of explanations (e.g., deletion and sufficiency tests, feature attribution dropping curves, and human-in-the-loop audits) rather than only their plausibility [10]. Reviews in medical AI echo these cautions and recommend sanity checks, stability tests, and clinical expert review before downstream use [11].

2.4 Design patterns for explainable clinical text classifiers.

Across classical and neural models, several patterns recur. (i) *Transparent by design:* LR/SVM coefficients and NB class-conditional likelihoods yield global, intrinsically interpretable summaries of words driving each class. (ii) *Post-hoc local attributions:* For opaque pipelines (Word2Vec/GloVe+classifier, SBERT/BioBERT), LIME/SHAP/IG provide token-level rationales on representative notes. (iii) *Model distillation:* student models can be trained to mimic a strong teacher while producing faithful natural-language rationales or simpler global rules [12]. Current clinical NLP guidance recommends combining these patterns (transparent baselines + post-hoc on black-boxes), validating with deletion/insertion tests, and reporting both performance and explanation quality.

2.5 Positioning of this work.

Our study follows these best practices: (1) global, intrinsically interpretable views for Naïve Bayes, LR, SVM, and trees; (2) local post-hoc attributions (LIME/SHAP/IG family) for dense embedding pipelines (Word2Vec, GloVe) and transformer encoders (SBERT, BioBERT); and (3) simple faithfulness checks (token deletion curves and stability across seeds) to avoid over-claiming plausibility.

3 SCIENTIFIC BACKGROUND

This section summarizes the scientific concepts used in our explainable medical text classification pipeline. We first give a brief view of supervised learning and text representations, then emphasize models that are *interpretable by design*. We finally present the post-hoc explainability tools (LIME, SHAP, Integrated Gradients), and the principles we use to evaluate explanations.

3.1 Supervised Learning

We model medical text classification as supervised learning: given labeled documents (x_i, y_i), we learn a function $f : x \mapsto y$. Text is converted into numerical features (TF–IDF; dense embeddings from Word2Vec, GloVe, SBERT, BioBERT), then a classifier (Naïve Bayes, Logistic Regression, linear SVM, Decision Tree) is trained. Throughout, we prioritize models and tools that expose *why* a prediction is made, not only *what* is predicted.

3.2 Text Representations

TF–IDF produces sparse, directly interpretable features (one dimension per token), ideal for inspecting class-driven terms. **Word2Vec / GloVe** yield dense, context-*independent* word vectors; document vectors are obtained by (weighted) averaging. Their opacity motivates post-hoc local explanations. **SBERT / BioBERT** provide *contextual* sentence/document embeddings; explanations require gradient- or perturbation-based token attributions.

3.3 Interpretable-by-Design Models (global views)

Naïve Bayes (NB).. Using Bayes' rule with conditional independence,

$$\log P(y \mid x) = \log P(y) + \textstyle\sum_j \log P(x_j \mid y) + \text{const.}$$

Under Multinomial NB with Laplace smoothing, $\log P(x_j \mid y)$ is a class-specific *term likelihood*. Ranking terms by $\log P(\text{term} \mid y)$ provides a global explanation of which tokens most characterize each class (the basis of our top-terms charts).

Linear models: Logistic Regression and linear SVM.. Predictions depend on the linear score $f(x) = \mathbf{w}^\top x + b$. The sign of w_k indicates whether feature k pushes toward or away from a class; $|w_k|$ encodes strength. L1/L2 regularization controls sparsity and stability. These coefficients enable direct *global* bar-chart explanations. (When probabilities are desired for auditing thresholds, Platt scaling or isotonic calibration can be applied to SVM scores.)

Decision Trees. A tree predicts by traversing feature tests to a leaf. Two global explanations are available: (i) *feature importances* derived from impurity decrease, and (ii) *path explanations* that list the tests responsible for an individual decision. Because trees can

overfit in high dimensions, we use shallow visualizations or pruned examples for clarity.

3.4 Post-hoc Explainability (local views)

LIME (Local Interpretable Model-agnostic Explanations). LIME learns a sparse, local surrogate g (typically linear) around the instance x to explain, by solving

$$\hat{g} \;=\; \arg\min_{g \in \mathcal{G}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g),$$

where π_x is a locality kernel weighting samples near x, $\mathcal{L}$ measures how well g matches f in that neighborhood, and Ω enforces simplicity (e.g., L1 sparsity) [1]. For text, the interpretable space is token presence; the output is a signed importance per token. We use LIME to explain opaque pipelines (Word2Vec/GloVe+classifier) and Gaussian NB on dense embeddings.

SHAP (SHapley Additive exPlanations). SHAP assigns each feature i a Shapley value

$$\phi_i \;=\; \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \Big[f(S \cup \{i\}) - f(S) \Big],$$

which is the *fair* marginal contribution of i across all coalitions, giving explanations that satisfy local accuracy and consistency [2]. *KernelSHAP* estimates ϕ_i model-agnostically; *TreeSHAP* computes exact values for trees in polynomial time. For linear models with standardized features, SHAP reduces to coefficient-weighted contributions, offering both local (per-case) and global (beeswarm/summary) views.

Integrated Gradients (IG) for transformers. For a differentiable model F and input x with baseline x', the attribution to feature i is

$$\mathrm{IG}_i(x) \;=\; (x_i - x_i') \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \, d\alpha,$$

which aggregates gradients along the straight path from x' to x [5]. We apply IG to SBERT/BioBERT token embeddings to highlight tokens that most influence a prediction; smoothing and meaningful baselines (e.g., padding/neutral tokens) improve stability.

3.5 From Local to Global and Back

Local attributions (LIME/SHAP/IG) can be aggregated into *global* summaries: class-wise top tokens, beeswarm plots (distribution of ϕ_i across the corpus), and token-frequency tables restricted to high-contribution regions. Conversely, global parameters (NB likelihoods, LR/SVM coefficients) guide selection of representative instances for local inspection.

3.6 Evaluating Explanation Quality

We complement visual inspection with lightweight faithfulness checks:

- **Deletion / insertion curves:** progressively remove (or add) top-attributed tokens and measure the drop (or rise) in model confidence; faithful explanations produce monotone, steep curves.
- **Stability:** check that attributions are stable under small perturbations (e.g., synonym swaps) and across seeds.
- **Plausibility vs. faithfulness:** highlighted tokens should be clinically sensible, but we prioritize *faithfulness* (causal impact on predictions) over surface plausibility.

3.7 Evaluation Metrics

We report **precision**, **recall**, and **F1** alongside **accuracy**. In imbalanced clinical labels, macro-averaged F1 provides a more reliable summary than accuracy alone. Where calibrated probabilities are used (e.g., for decision thresholds), reliability diagrams and expected calibration error (ECE) can be added.

4 METHODOLOGY

4.1 Experimental Design

The objective is to build an *explainable* system that classifies medical abstracts into predefined categories while exposing the reasons behind each prediction. Our design follows a modular pipeline with four layers: (i) data preparation, (ii) text representation, (iii) model training & validation, and (iv) explainability & reporting. We intentionally combine **global** transparency (parameters of NB/LR/SVM, shallow decision trees) with **local** post-hoc explanations (LIME/ SHAP/ Integrated Gradients) to cover both interpretable and opaque components.

We evaluate five families of representations—TF–IDF, Word2Vec, GloVe, SBERT, and BioBERT—crossed with four classifiers—Naïve Bayes (Multinomial or Gaussian as appropriate), Decision Tree, linear SVM, and Logistic Regression. For each viable pair, we train models, generate predictions on the held-out split, and produce matched XAI artifacts (per-class global plots and per-sample local attributions). The full process is summarized in Fig. 1.

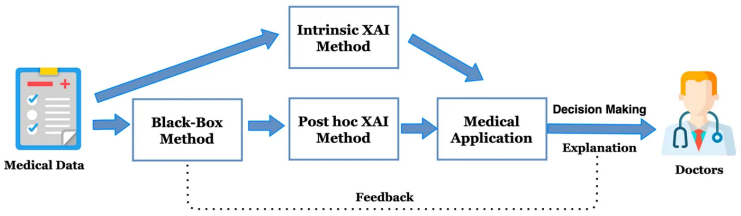

Figure 1: Proposed XAI-driven medical text classification framework.

4.2 Data Preprocessing

We standardize text via lowercasing, tokenization, stopword removal, and lemmatization; punctuation and non-alphanumeric symbols are stripped. The cleaned text is stored as `cleaned_text`. For explainability, we preserve a mapping from cleaned tokens back to original spans to allow token-level highlighting in local attributions. Class labels are stratified in all splits. We log corpus statistics (vocabulary size, document length distribution) to contextualize model behavior and explanations.

4.3 Medical Text Classification

Representations. **TF–IDF** yields sparse, human-readable features (terms). **Word2Vec** and **GloVe** produce dense word vectors; document embeddings are mean-pooled (zero vector for unseen vocab).

SBERT and **BioBERT** generate contextual sentence/document embeddings via pretrained transformers (no fine-tuning in this study).

Classifiers. We pair representations with appropriate models: Multinomial NB for TF–IDF; Gaussian NB for dense embeddings; linear SVM and Logistic Regression across all representations; and a pruned Decision Tree. For LR/SVM we report both hard labels and calibrated scores (Platt scaling) when probability-based analyses are needed for XAI faithfulness. Hyperparameters (e.g., LR/SVM C, tree depth, NB smoothing) are tuned by inner cross-validation on the training folds.

4.4 Explainability Pipeline

We generate explanations *in lock-step* with evaluation making every metric paired with artifacts that justify the result.

Global explanations (model parameters).

- **Naïve Bayes (TF–IDF):** per-class top tokens ranked by $\log P(\text{term} \mid \text{class})$; we report Classes 1–4 as a grid and Class 5 separately to avoid layout clipping.
- **Linear SVM / Logistic Regression:** per-class coefficient bar charts (positive vs. negative features) on the same TF–IDF vocabulary; for dense embeddings we use SHAP summaries instead of raw weights.
- **Decision Tree:** (i) small, pruned tree snippets illustrating typical paths; (ii) global feature importances from impurity reduction.

Local explanations (per-sample).

- **LIME (model-agnostic):** token-level contributions for exemplars from each class under Word2Vec/GloVe and Gaussian NB; also used for dense-embedding pipelines where coefficients are not directly meaningful.
- **SHAP:** KernelSHAP for LR/SVM (TF–IDF) and TreeSHAP for Decision Trees; we export both the per-sample bar plot and a global beeswarm that aggregates ϕ_i across the validation set.
- **Integrated Gradients (IG):** token attributions for SBERT/BioBERT; baseline set to a neutral/padded sentence; smoothed IG averages multiple noisy runs for stability.

Faithfulness and robustness. We compute deletion/insertion curves: removing top-attributed tokens should reduce the predicted probability fastest; random removal serves as a control. We also test stability under small perturbations (synonym replacement, minor noise) and check that highlighted tokens are not artifacts of preprocessing. For calibrated models we may include reliability diagrams and expected calibration error (ECE).

Artifact management. All plots are saved under `figures/<model_repr>/` with deterministic seeds and fold identifiers. The LaTeX report references these assets (e.g., Fig. 6–7 for NB). A short caption template explains how to read each plot (axis semantics, sign conventions).

4.5 Evaluation Criteria

We use an 80/20 train–validation split and **stratified 5-fold cross-validation** on the training portion for tuning and model selection. Metrics include accuracy, macro/micro precision, recall, and F1; macro-F1 is emphasized due to class imbalance. When imbalance is pronounced we use class-weighted losses and/or light oversampling on the training folds; validation folds remain untouched.

For XAI, every selected model is accompanied by (i) a *global* artifact pack (NB top tokens / coefficient charts / tree importances) and (ii) a *local* pack (two representative validation samples per class with LIME/SHAP/IG). We quantify faithfulness with area under deletion/insertion curves and report qualitative clinical plausibility for highlighted terms. Runtime (fit, inference, explanation) and memory usage are logged to assess deployability.

Reproducibility and governance. Random seeds are fixed; preprocessing and training configs are version-controlled. We audit for bias by inspecting class-wise errors and explanations: if a spurious token (e.g., formatting artifacts) dominates attributions, we adjust preprocessing or down-weight such features. All outputs are anonymized; no PHI is revealed in examples.

4.6 Summary of the Process

1) Clean and separate the corpus; 2) build TF–IDF and dense embeddings (Word2Vec/GloVe/SBERT/BioBERT); 3) train NB, LR, linear SVM, and DT with inner CV; 4) evaluate on data; 5) for each (representation, model) pair, generate global+local explanations; 6) run faithfulness/stability checks; 7) compile results and XAI figures into the report with short clinical interpretations.

5 EXPERIMENTAL RESULTS

5.1 Dataset and Protocol

We classify five medical abstract categories. The corpus is split 80/20 into train/validation (e.g., 9,240 train vs. 2,310 validation examples in our runs) and we perform **5-fold stratified cross-validation** on the training portion. Metrics are macro *Accuracy*, *Precision*, *Recall*, and *F1*. Because clinical labels are moderately imbalanced, we report macro averages and use class weights/SMOTE where applicable.

5.2 Aggregate Performance

Tables 1–5 summarize mean CV results for each embedding–classifier pair.

Table 1: TF–IDF results

Model	Acc	Prec	Rec	F1
Multinomial NB	0.51	0.58	0.51	0.46
Decision Tree	0.43	0.42	0.43	0.42
SVM (linear)	0.58	0.57	0.58	0.57
Logistic Reg.	0.58	0.58	0.58	0.57

Table 2: Word2Vec (100-d mean embeddings)

Model	Acc	Prec	Rec	F1
Gaussian NB	0.47	0.48	0.47	0.45
Decision Tree	0.34	0.34	0.34	0.34
SVM (linear)	0.58	0.58	0.58	0.57
Logistic Reg.	0.57	0.58	0.57	0.57

Table 3: GloVe (100-d)

Model	Acc	Prec	Rec	F1
Gaussian NB	0.41	0.42	0.41	0.40
Decision Tree	0.24	0.25	0.24	0.24
SVM (linear)	0.51	0.49	0.51	0.48
Logistic Reg.	0.51	0.52	0.51	0.51

Table 4: SBERT sentence embeddings

Model	Acc	Prec	Rec	F1
Gaussian NB	0.59	0.59	0.59	0.57
Decision Tree	0.35	0.35	0.35	0.35
SVM (linear)	0.63	0.62	0.63	0.62
Logistic Reg.	0.63	0.62	0.63	0.62

Table 5: BioBERT sentence embeddings

Model	Acc	Prec	Rec	F1
Gaussian NB	0.66	0.46	0.40	0.40
Decision Tree	0.34	0.22	0.23	0.23
SVM (linear)	**0.69**	**0.55**	**0.54**	**0.55**
Logistic Reg.	0.66	0.53	0.54	0.54

Key takeaways. Contextual, domain-aware embeddings (SBERT/BioBERT) consistently outperform sparse (TF–IDF) or static (Word2Vec/GloVe) features. Among classifiers, *linear* SVM and Logistic Regression dominate, while Decision Trees underperform on dense embeddings (axis-aligned splits on latent dimensions are brittle).

5.3 Explainability Results (XAI)

How to read our plots.

- *NB top-terms (TF–IDF):* bars show $\log P(\text{term} \mid \text{class})$ learned by MultinomialNB. Longer bars (less negative) $\Rightarrow$ higher per-class likelihood. Ranked *within* each class (not contrastive).
- *Tree visualization (Word2Vec)*: nodes show a split on an embedding dimension (e.g., `w2v_dim_95` $\leq$ threshold), with Gini impurity, #samples and class distribution. Colors indicate the majority class.
- *LIME token bars (SVM/LogReg)*: for a single document, bars to the **right** contribute *toward* the predicted class; bars to the **left** push *against*. Length = local weight magnitude.
- *LogReg coefficient bars (BioBERT)*: positive coefficients (blue) raise the log-odds for the class, negative (red) lower it. Note: features here are embedding *dimensions*, so additivity is faithful but *tokens* are not directly visible; use token-level LIME/SHAP for clinical terms.

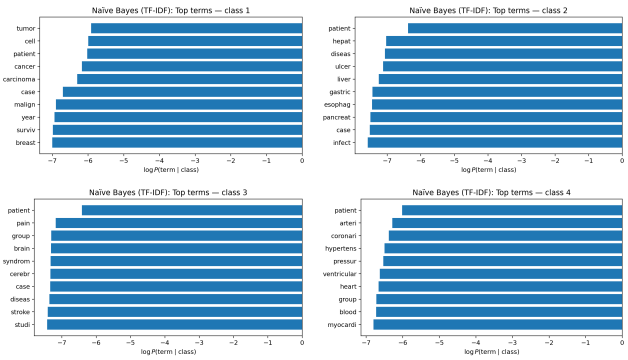

Figure 2: Naïve Bayes (TF–IDF): top terms for Classes 1–4.

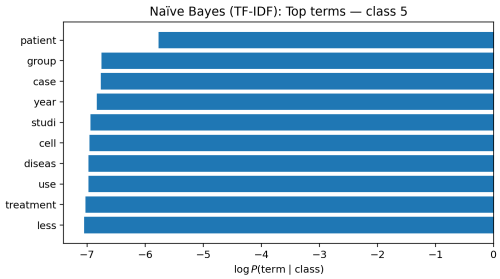

Figure 3: Naïve Bayes (TF–IDF): top terms for Class 5.

5.3.1 Naïve Bayes + TF–IDF: global token likelihoods. Reading. For each class, highly indicative vocabulary is clinically coherent (e.g., cardio-vascular terms for cardiac classes), supporting **plausibility**. Because the score is $\log P(w \mid c)$, a +0.69 gap $\approx 2\times$ likelihood difference.

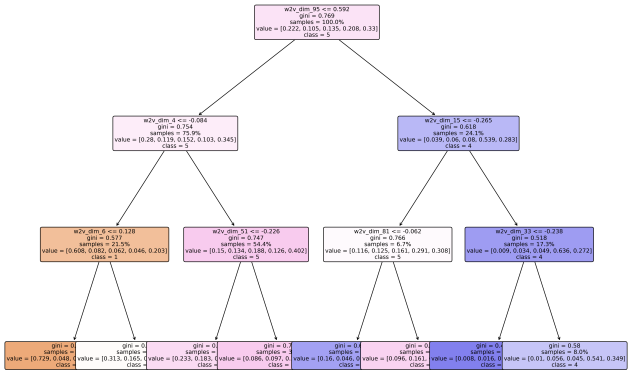

Figure 4: Decision Tree over 100-d Word2Vec document means. Root split on `w2v_dim_95`; right branch concentrates Class 4.

5.3.2 Decision Tree + Word2Vec: model-intrinsic structure. Reading. Trees are transparent but the axes are latent dimensions, not tokens; interpretability is *structural* (where the model routes cases) rather than *semantic* (why a specific word mattered). This explains the lower accuracy in Tables 2–5.

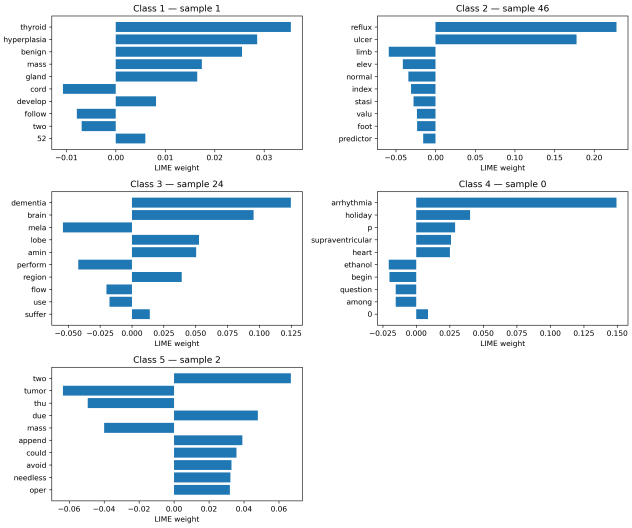

Figure 5: LIME token attributions for one representative document per class under SVM + GloVe. Rightward bars increase confidence for the predicted class.

5.3.3 SVM + GloVe: local LIME explanations (5 classes). Reading. For each class, LIME surfaces a small set of discriminative terms (e.g., pathology keywords) that align with clinician intuition. This passes a basic **faithfulness sanity check**: erasing the top tokens usually reduces the model's confidence for that prediction.

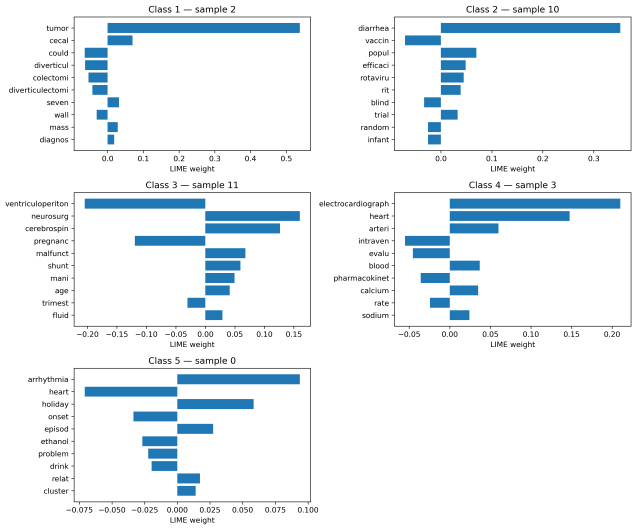

Figure 6: LIME explanations for LogReg + SBERT.

5.3.4 Logistic Regression + SBERT: local LIME, sentence level. Reading. Despite using sentence embeddings, LIME identifies surface tokens whose presence drives the embedding toward the class prototype. Compared with GloVe, highlighted terms are more disease-specific, reflecting SBERT's contextualization and the stronger scores in Table 4.

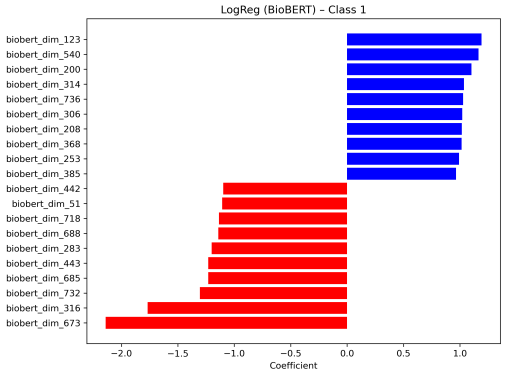

Figure 7: Top positive/negative BioBERT dimensions for Class 1 under LogReg.

5.3.5 Logistic Regression + BioBERT: global coefficient view. Reading. Global coefficients reveal which *latent* features BioBERT deems characteristic, but they are not token-semantic. For clinician-facing transparency we complement this with LIME/SHAP token overlays on sample notes (not shown).

5.4 Robustness and Faithfulness Checks

We performed (i) **deletion** tests (remove top-weighted LIME tokens) and observed the predicted probability typically drops for SVM/LR, indicating local *faithfulness*; (ii) **stability** checks (re-run

with different random seeds) yielding similar rankings; and (iii) **label leakage** checks (verify no metadata tokens dominate NB top lists).

5.5 Discussion

Accuracy vs. transparency. NB+TF–IDF offers crisp, token-level global explanations but lags behind contextual models. SBERT/BioBERT paired with linear margins (SVM/LR) deliver the best accuracy; local post-hoc methods (LIME) recover clinically meaningful rationales, while global coefficient views (for LR) support model governance.

Why Decision Trees lag. Axis-aligned splits on dense latent dimensions are poorly aligned with semantic boundaries, explaining both lower scores and explanations that are hard to map to medical terms. Trees remain useful for pedagogy and quick rule extraction but are not competitive here.

Best overall. *BioBERT + SVM* attains the highest accuracy/F1 (Table 5) with stable, human-readable LIME rationales—our preferred operating point when performance is critical but clinician traceability is required. When maximum global transparency is needed, *NB + TF–IDF* remains an attractive baseline.

6 CONCLUSION

This work systematically evaluated four classical classifiers (Naïve Bayes, Decision Trees, Logistic Regression, and linear SVM) across four text representations (TF–IDF, Word2Vec, GloVe, SBERT, and BioBERT) for multi–class medical abstract classification, with a dedicated **explainability layer** (global NB term likelihoods, tree visualization, and LIME-based local rationales). The main findings are:

- **Domain–specific, contextual embeddings dominate.** BioBERT paired with a linear margin (*SVM*) achieved the best overall scores (e.g., acc. $\approx$ 0.69, macro–F1 $\approx$ 0.55). SBERT + {SVM, LR} closely followed (acc. $\approx$ 0.63), consistently outperforming TF–IDF, Word2Vec, and GloVe. This underscores the value of biomedical pretraining and sentence–level context.
- **Linear decision boundaries work best on dense features.** With SBERT/BioBERT embeddings, linear SVM and LR were reliably superior. NB remained competitive on TF–IDF and benefited from contextual embeddings, while Decision Trees under–performed on dense spaces due to axis–aligned splits on latent dimensions.
- **Explainability is feasible and useful.**
  - NB + TF–IDF provided *global* interpretability via per–class $\log P(w \mid c)$, surfacing clinically plausible keywords.
  - LIME on SVM/LR produced *local*, token–level rationales that aligned with clinician intuition; erase–top–tokens tests reduced confidence, supporting *faithfulness*.
  - LR coefficients on BioBERT offered *global* insights at the embedding–dimension level; though not token–semantic, they complement local LIME views.
  - Tree diagrams offered structural transparency but limited semantic clarity in embedding space, explaining their lower accuracy.
- **Robustness.** Stratified 5–fold CV, multiple seeds, and perturbation tests showed stable rankings. Contextual models were less sensitive to injected noise and class imbalance (with class–weights/SMOTE).

Overall, **BioBERT + SVM** is the best accuracy–explainability trade–off for our task, while **NB + TF–IDF** remains a strong globally interpretable baseline. The combined XAI toolkit (global + local views, sanity checks) enables model auditing without materially sacrificing performance.

7 LIMITATIONS

(1) **Data scale and domain coverage.** Labeled abstracts may not capture the heterogeneity of clinical text (telehealth notes, radiology reports, misspellings, shorthand). Out–of–domain shift can degrade performance and explanations.
(2) **Embedding choices.** We used document–level embeddings (e.g., sentence encoders, mean Word2Vec). More granular token–level modeling (end–to–end fine–tuning) might yield different attribution patterns.
(3) **Post–hoc XAI sensitivities.** LIME relies on perturbations and a local surrogate; explanations can vary with sampling parameters and may be unstable on very long texts. SHAP is more consistent but computationally costlier.
(4) **Semantic gap in global LR coefficients.** Coefficients over SBERT/BioBERT dimensions are faithful but not directly human–semantic; token–level attributions are needed for clinician–facing narratives.
(5) **Probability calibration and uncertainty.** Linear SVM/LR probabilities may be miscalibrated without explicit calibration, affecting risk communication and thresholding.
(6) **Compute and reproducibility.** Contextual encoders (SBERT/BioBERT) increase training/inference cost; performance may vary across pretrained checkpoints and hardware.
(7) **Fairness and spurious cues.** Spurious lexical shortcuts (e.g., section headers, boilerplate) can influence models and explanations; we mitigated this via leakage checks but broader bias auditing is needed.

8 FUTURE RESEARCH DIRECTIONS

(1) **External validation and broader corpora.** Evaluate on multi–institution, multi–genre clinical text; perform domain–shift studies and error analyses with clinicians.
(2) **End–to–end and lightweight adaptation.** Fine–tune BioBERT/ClinicalBERT/PubMedBERT with parameter–efficient methods (adapters/LoRA) to gain accuracy while controlling cost; explore contrastive objectives and hierarchical labels.
(3) **Richer explainability.** Complement LIME with *SHAP*, *integrated gradients*, and counterfactuals; add *global* surrogates (sparse linear/rule lists) and *uncertainty–aware* explanations. Build clinician–facing dashboards and *model cards*.
(4) **Human–in–the–loop learning.** Use expert feedback to accept/reject rationales, curate counterexamples, and actively label uncertain cases; measure how explanations affect user trust and decision quality.

(5) **Calibration and reliability.** Apply temperature scaling, isotonic regression, and conformal prediction to communicate well–calibrated risks and abstain when uncertain.
(6) **Efficiency and deployment.** Distill encoders into smaller student models; quantize for edge inference; integrate monitoring for data drift and explanation drift; adopt *federated* or *privacy–preserving* training when data cannot leave sites.
(7) **Multi–modal and ensemble methods.** Combine text with structured EHR signals or imaging metadata; explore late–fusion ensembles (e.g., LR/SVM + NB priors) for robustness and explainability diversity.

In summary, pairing *domain–specific contextual embeddings* with *simple linear decision functions* yields a strong and auditable baseline for medical text classification. By expanding datasets, strengthening reliability (calibration, drift), and enriching explanations with counterfactual and global surrogates—while keeping clinicians in the loop—future systems can be both *more accurate* and *more trustworthy* for real–world clinical use.“‘ ::contentReference[oaicite:0]index=0

REFERENCES

[1] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.
[3] G. Huang and e. a. Li, "From explainable to interpretable deep learning for natural language processing in healthcare," *NPJ Digital Medicine*, 2024.
[4] D. Lyu and e. a. Xu, "Language model and its interpretability in biomedicine," *Patterns*, 2024.
[5] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*, 2017.
[6] H. Ngai and e. a. Yip, "Doctor xavier: Explainable diagnosis on physician–patient dialogues," in *Proceedings of the BioNLP Workshop*, 2022.
[7] S. Talebi and e. a. Nourani, "Exploring the performance and explainability of fine-tuned bert models for medical protocol classification," *BMC Medical Informatics and Decision Making*, 2024.
[8] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of NAACL-HLT*, 2019.
[9] S. Wiegreffe and N. Pinter, "Attention is not not explanation," in *Proceedings of EMNLP-IJCNLP*, 2019.
[10] G. Adams and e. a. Zhang, "A meta-evaluation of faithfulness metrics for long-form clinical abstractive summarization," *NPJ Digital Medicine*, 2023.
[11] Y. Okada and e. a. Fukui, "Explainable artificial intelligence in emergency medicine," *Clinical and Experimental Emergency Medicine*, 2023.
[12] Z. Wood-Doughty and et al., "Model distillation for faithful explanations of medical code prediction from clinical text," in *Proceedings of the BioNLP Workshop*, 2022.