Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 14 July 2026
Revised 21 Aug 2026
Accepted 04 Sep 2026

Natural Resources for Human Health 2026, Vol. 6 (19s): 160–175

KEYWORDS:
large language models; synthetic data; few-shot prompting; clinical NLP; dermatology; text classification

Derma Synth: GPT-4-Based Synthetic Text Generation for Dermatology via Meta-Prompting and Few-Shot Conditioning

Divya Mishra[1]*, Amit Agrawal[2], Ankush Mittal[3], Deepak Painuli[4]

[134]Computer Science and Engineering, COER University, 7 km Roorkee–Haridwar Highway, Roorkee 247667, Uttarakhand, India.
[2]AI Research Scientist, Wells Fargo, Bangalore, Karnataka, India.
*Corresponding author: divya19aug@gmail.com.
Co-authors: amitagrawal1909@gmail.com; dr.ankush.mittal@gmail.com; prof.deepak.painuli@gmail.com

ABSTRACT: Automated classification of dermatological conditions from free text is limited by the scarcity of labelled clinical and patient-narrative corpora. We present DermaSynth, a meta-prompting framework that uses GPT-4 to expand a seed set of 67 real dermatology texts into a synthetic dataset of 653 labelled texts covering seven conditions (acne, contact dermatitis, eczema, folliculitis, psoriasis, scabies and vitiligo; 82–100 texts per class). Generation combines a dermatologist-persona context prompt, a diversity prompt, and in-context exemplars drawn from the seed set ($k \in \{2, 4, 6, 8\}$). We evaluate the dataset with Random Forest (RF) and Multinomial Naïve Bayes (MNB) classifiers under repeated stratified cross-validation, and through ratings by three board-certified dermatologists. RF reaches 94.55 ± 0.55% accuracy (macro-F1 0.944) and MNB 91.58 ± 0.45% (macro-F1 0.915) in 5×5 cross-validation, and the RF advantage is statistically significant on paired predictions (McNemar $p = 0.003$). However, 58.7% of the texts name their own disease, and masking the names lowers accuracy to 84.6% (RF) and 83.5% (MNB), so about 8–10 points of accuracy come from keyword cues. Texts also have a low length (mean 21.7 words), while contact dermatitis is very templated. Data was evaluated 4.53/5 by dermatologists on four different dimensions (realism 4.3/5). All classification results are on synthetic text; transfer to real patient text remains to be tested.

1. INTRODUCTION

Skin diseases are among the most prevalent human health conditions [1] and carry a measurable psychological burden [2], which makes timely and accurate triage valuable. Most automated dermatology research has been built on image corpora, such as the dermatologist-level classifier of Esteva et al. [3] and the HAM10000 dermatoscopic collection [4]. Such resources depend on expert annotation and specialised imaging equipment and are expensive to extend.

Patient-written symptom descriptions are a cheaper and more accessible signal and underpin symptom checkers, triage chatbots and NLP-based decision support. However, to our knowledge, large annotated corpora of dermatological text are not publicly available. The seed set used in this work contains only 67 labelled texts [5]. This scarcity limits supervised training of text classifiers and motivates data augmentation.

Large language models (LLMs) follow instructions and adapt to new tasks from a few in-context examples [6, 7], and have been used to generate synthetic clinical text in several settings [8, 9, 10]. Generation through synthetics also allows control over the class distribution and style. However, the utility of LLMs for the generation of training data depends on the task [11], and recursive training using synthesized data can harm models [12]. The use of synthetic data thus requires a rigorous assessment.

Table 1 shows real seed examples that illustrate the variation in register (clinical description versus first-person narrative), severity and vocabulary. Table 2 shows samples produced by DermaSynth from such seeds.

Table 1. Examples from the real seed set.

Disease	Text
Acne	Acne manifests in various forms, showcasing inflammatory symptoms such as papules, pustules, nodules, and cysts, each presenting different sizes and levels of tenderness. In addition, non-inflammatory signs such as blackheads and whiteheads may appear, accompanied by redness, swelling, and oily skin. The condition often leads to scarring, especially when severe or untreated, impacting the patient's emotional well-being.
Contact dermatitis	My hands are cracked and my skin is flaking off in certain areas. Do you think the chemicals and oils that I come into contact with at work could be the cause?
Eczema	Eczema is otherwise known as atopic dermatitis and causes diverse symptoms depending on the look and feel of the skin. Some of the common symptoms include having red patches of skin, which are normally accompanied by unbearable itching, making the person scratch and causing the skin to thicken.
Folliculitis	I have small reddish lumps on my back and chest and these lumps have a feel as if I got razor bumps.

Table 2. Examples from the synthetic dataset generated with GPT-4.

Disease	Text
Acne	I have very oily skin and no matter how many times I cleanse my face, it is always oily. My breakouts are becoming more frequent and are mostly on my nose and forehead. Is there any product that can control oil production without drying my skin?
Contact dermatitis	My hands have been itchy and red for quite some time. This is following my use of a different kind of soap in my place of work.
Eczema	My child's eczema seems to flare up when the weather changes. I am trying to figure out how to manage it without using too many steroids.
Folliculitis	I have itchy, pus-filled bumps on my armpits after starting a new deodorant. They are uncomfortable and make me self-conscious.

The contributions of this paper are as follows.

- **Generation framework.** A GPT-4-based meta-prompting framework that combines a context prompt, a diversity prompt and k-shot exemplars, specified through explicit prompt templates and a generation algorithm (Section 4).
- **Synthetic dataset.** A set of 653 synthetic texts (82–100 per class) for seven dermatological conditions, produced from 67 seed texts.
- **Statistically reported extrinsic evaluation.** RF and MNB classifiers evaluated with repeated stratified cross-validation, paired tests, per-class precision, recall, F1 and AUC, confusion analysis, Wilson intervals, Cohen's κ and the multiclass Matthews correlation coefficient (Section 5).

- **Expert evaluation.** A rating study by three board-certified dermatologists on coherence, symptom accuracy, diversity and realism.
- **Validity analysis.** A keyword-leakage ablation, split-variability analysis, error analysis and learning curve.

The remainder of the paper is organised as follows. Section 2 reviews related work, Section 3 formalises the problem, Section 4 describes the method and experimental setup, Section 5 reports results, Section 6 discusses limitations, and Section 7 concludes.

2. RELATED WORK

2.1 Image-based dermatological classification

Esteva et al. trained a convolutional network on 129,450 clinical images and reported dermatologist-level performance on skin-cancer classification tasks [3]. HAM10000 provides 10,015 dermatoscopic images across seven diagnostic categories [4]. These resources show what large expert-labelled image sets enable, but their construction relies on annotation effort and specialised acquisition equipment. Text-based alternatives are comparatively less developed.

2.2 Clinical NLP and few-shot LLMs

Rule-based NLP systems for encoding clinical documents date back to the early 2000s [13]. More recently, LLMs have been applied to clinical text with few or no task-specific labels. GPT-3 demonstrated in-context few-shot learning [6], and GPT-4 extended this capability [7]. Agrawal et al. showed that LLMs act as few-shot clinical information extractors [14], and Labrak et al. compared zero- and few-shot behaviour of instruction-tuned models across clinical and biomedical tasks [15]. These works motivate using a small number of real exemplars to condition generation.

2.3 Synthetic clinical text

GatorTronGPT generated synthetic clinical text that was then used to train downstream models [8]. Asclepius built a publicly shareable clinical LLM from synthetic clinical notes [9]. MedSyn combined LLMs with a medical knowledge graph to generate over 41,000 Russian-language clinical notes covering 219 ICD-10 codes and reported accuracy gains of up to 17.8% on challenging ICD codes [10]. Belkadi et al. generated synthetic free-text records with a masked language model and measured a re-identification risk of 3.5% [16]. The self-instruct paradigm similarly bootstraps training data from a model's own generations starting from a small seed set [17]. In medical imaging, GAN-based synthetic augmentation improved liver-lesion classification [18].

Synthetic data is not uniformly beneficial. Li et al. found that the value of LLM-generated data for text classification depends on the task and decreases for more subjective tasks [11], and Shumailov et al. showed that training on recursively generated data leads to model collapse [12]. These findings motivate reporting uncertainty, per-class behaviour and limitations alongside headline accuracy.

2.4 Positioning of this work

Prior synthetic-text work targets clinical notes, ICD coding or general medical text. To our knowledge, DermaSynth is the first framework focused on patient-facing dermatological text with a seed set of fewer than 100 examples. Unlike MedSyn [10], it uses no knowledge graph and relies on prompt design and exemplars alone. Unlike MLM-based approaches [16], it treats the generator as a black box.

3. PROBLEM FORMULATION

Let $\mathcal{X}$ be the space of dermatological texts and $\mathcal{Y} = \{y_1, \ldots, y_K\}$ the label set, with K = 7 classes: acne, contact dermatitis, eczema, folliculitis, psoriasis, scabies and vitiligo. Given a seed set $D_{real} = \{(x_i, y_i)\}_{i=1}^{N}$ with N = 67, the aim is to construct a synthetic set $D_{syn} = \{(x'_j, y'_j)\}_{j=1}^{M}$ with M = 653 (M ≈ 9.7 N) by sampling from an LLM G conditioned on a target label, a few exemplars and a set of natural-language constraints:

$$x'_j \sim P_G(x \mid y'_j, D_k, c) \quad (1)$$

Here y'_j is the target label, $D_k \subset D_{real}$ is a set of k exemplars with $k \in \{2, 4, 6, 8\}$, and c collects the constraints (dermatologist persona, diversity instruction, approximate length). GPT-4 is accessed as a black box, so P_G is not available in closed form and cannot be optimised by gradient descent. The criteria below are therefore used as quality-control and model-selection criteria that guide prompt refinement and filtering, not as training losses.

3.1 Quality criteria

Lexical diversity. Texts for a class should not be near-copies of each other. We measure redundancy by mean self-BLEU (lower is more diverse):

$$L_{\text{div}} = \frac{1}{M} \sum_{j=1}^{M} \text{BLEU}_4\left(x'_j,\ D_{\text{syn}} \setminus \{x'_j\}\right) \quad (2)$$

Domain alignment. Synthetic texts should lie close to real texts in a sentence-embedding space φ. With the mean embedding μ_S of a text set S,

$$L_{\text{align}} = 1 - \cos\left(\mu_{\text{syn}},\ \mu_{\text{real}}\right), \qquad \mu_S = \frac{1}{|S|} \sum_{x \in S} \varphi(x) \quad (3)$$

Alignment of the means is necessary but not sufficient, so the maximum cosine similarity of each synthetic text to any seed text is a further useful check, because it exposes near-verbatim copying of the seeds.

Classification utility. A classifier trained on D_{syn} and tested on real text (TSTR) should perform like one trained and tested on real text (TRTR):

$$L_{\text{util}} = A_{\text{TRTR}} - A_{\text{TSTR}} \quad (4)$$

If a single scalar summary is needed, the criteria can be combined as $L_{total} = \lambda_1 L_{div} + \lambda_2 L_{align} + \lambda_3 L_{util}$ with $\lambda_i > 0$. In this work the criteria guide prompt iteration and filtering and are not optimised jointly.

4. Methodology

The pipeline process is described in Figure 1 below. The seed texts and the prompts are assembled using the prompt assembler. The GPT-4 synthesizes a text, which then goes through the filter, eliminating repetitions and unrealistic answers. nual reviews are conducted periodically to improve the prompt. Finally, the obtained dataset Dsyn undergoes external evaluation, expert assessment, and intrinsic checks.

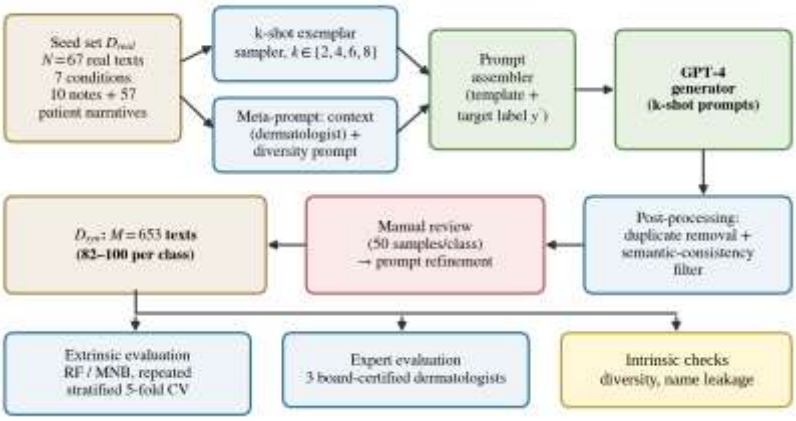


Fig. 1. The DermaSynth pipeline: seeding sets, meta-prompt parts, GPT-4 text generation, filtering with human assessment, and evaluation (classification in Section 5.2, intrinsic and leakage analysis in Sections 5.1 and 5.5, expert assessments in Section 5.7).

4.1 Seed data

The seed set Dreal consists of N = 67 texts for the seven classes (9-11 texts per class), which come from a publicly available Kaggle dataset [5]. This seed set includes 10 formal clinical texts and 57 informal patient narratives containing 50-150 words each (Table 1). As 67 texts are not sufficient to train a reliable classifier, Dreal is used only for conditioning the generator.

4.2 Meta-prompt design

The prompts contain three aspects (Listing 1). The first prompt is called context prompt that provides the model with a dermatologist's point of view so that the vocabulary and symptoms become more realistic. The second prompt is called the diversity prompt where the model is asked to change body location, severity, time span, trigger, and register. The third aspect is the few-shot block.

Listing 1. Meta-prompt template

[CONTEXT PROMPT]

You are a board-certified dermatologist as well as someone who understands how patients talk about their skin conditions.

[DIVERSITY PROMPT]

Change the site, level, duration, trigger and the style of presentation (clinical language or personal statement). Avoid repetition of any previous text.

[FEW-SHOT BLOCK]

In view of the following concrete examples: [x_1, y_1]: "<text 1>", [x_2, y_2]: "<text 2>", …, [x_k, y_k]: "<text k>", construct a 100-word story of the patient suffering from <target disease y'> that follows these patterns and symptomatology.

4.3 Few-shot conditioning

Generation was run under four conditioning levels, k ∈ {2, 4, 6, 8}. For each query, k exemplars were randomly selected from Dreal. These were not constrained by the target label, meaning that, for instance, the prompt for folliculitis can include an eczema description and a psoriasis story. The exemplars thus carry mostly information regarding register and style, while the target label from the instruction carries the disease information. All outputs corresponding to different k values are aggregated into Dsyn. There was no separate ablation done per k and is left for future work.

4.4 Generation loop and quality control

Algorithm 1. DermaSynth generation

Input: seed set D_{real}; label set $\mathcal{Y}$; k-set K = {2, 4, 6, 8}; per-class target $n_c \approx 100$; prompts p_{ctx}, p_{div}

Output: synthetic set D_{syn} (M = 653 texts after filtering)

$D_{syn} \leftarrow \emptyset$
for each class $y \in \mathcal{Y}$ do
  while $|D_{syn}(y)| < n_c$ do
    sample k from K; sample $D_k \subset D_{real}$ uniformly at random, $|D_k| = k$
    prompt ← Assemble(p_{ctx}, p_{div}, D_k, y)
    x' ← GPT-4(prompt)
    if Filter(x', D_{syn}) accepts then $D_{syn} \leftarrow D_{syn} \cup \{(x', y)\}$
    every 50 accepted samples: manual review against D_{real}; refine p_{ctx}, p_{div} if needed
return D_{syn}

Prompts were tuned through an iterative process. For every 50 prompts produced by the model for each class, some prompts were manually inspected for their coherence and changes were made if there were any issues (e.g., "focus on dermatological symptoms" was added in case there were unrelated topics). Duplicate and implausible data was filtered out by post-processing the data using a medical BERT model.

4.5 Experimental setup

The classification experiment uses scikit-learn [19]. The algorithms used for classification experiments are Random Forest [20] and Multinomial Naïve Bayes [21]. The configurations used are shown in Table 3. Protocol A is a single random split (not stratified) that mirrors the 196-test texts used in an earlier iteration of this research. Because one split of this size is unstable, the main results use protocol C, in which every text is classified out-of-fold in each of five repeats; protocol B quantifies split-to-split variation. Protocol D replaces each disease name with a placeholder token to measure keyword reliance.

Table 3. Experimental configuration.

Item	Setting
Seed set	N = 67 real texts, 7 classes, 9–11 texts per class [5]
Generator	GPT-4 [7]
Conditioning	$k \in \{2, 4, 6, 8\}$ random exemplars from the seed set; pooled
Generated set	M = 653 texts (82–100 per class; mean 21.7 words, range 8–74) after filtering
Features	Random Forest: TF-IDF (unigram); Multinomial NB: raw token counts. Vectorizers are fitted on training folds only
Classifiers	scikit-learn [19]: Random Forest with 200 trees (other settings default); Multinomial NB with $\alpha = 1$
Protocols	(A) one random 70/30 split, random_state = 42 (457 train, 196 test); (B) 20 random 70/30 splits; (C) 5×5 repeated stratified 5-fold cross-validation; (D) disease names masked, repeat C; (E) learning curve
Metrics	Accuracy, precision, recall, F1 (per class, macro, weighted), one-vs-rest AUC, Cohen's $\varkappa$, MCC, Wilson 95% intervals, exact McNemar test

4.6 Evaluation metrics

For class c, precision P_c, recall R_c and F1-score are defined as follows:

$$F1_c = \frac{2P_cR_c}{P_c+R_c}, \qquad P_c = \frac{TP_c}{TP_c+FP_c}, \qquad R_c = \frac{TP_c}{TP_c+FN_c} \tag{5}$$

Macro averages weight all classes equally and weighted averages weight classes by test support. Because the test classes are not exactly balanced, both are reported. Accuracy confidence intervals use the Wilson score interval [22], which behaves well for proportions near one:

$$\tilde{p}_{\pm} = \frac{\hat{p}+\frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2}{4n^2}}}{1+\frac{z^2}{n}} \tag{6}$$

Where $\hat{p}$ represents the estimated accuracy, n is the total number of test texts and z = 1.96. The comparison of two classification algorithms tested using the same test texts is done using the exact McNemar test [23] where the texts that one classifier made a mistake but not the other are used. Cohen's kappa [24] and the multi-class Matthews correlation coefficient [25] consider random agreement. AUC values are one-vs-rest.

4.7 Expert evaluation protocol

Three board-certified dermatologists evaluated random selections of synthesized texts and reference texts for four attributes: medical plausibility, symptom accuracy, variety of patient symptoms, and realism, where each attribute was rated from 1 to 5.

5. RESULTS AND DISCUSSION

Classification results are reported below based on the final dataset comprising 653 texts. The results are obtained through the corresponding analysis script run on the dataset file; the protocols are explained in Section 4.5.

5.1 Dataset characteristics

The dataset comprises 653 texts, each having between 82 (vitiligo) and 100 (psoriasis) texts per class (Table 4). The texts are quite short, consisting of an average of 21.7 words per text (minimum 8 and maximum 74), and 52 texts consist of less than 15 words. The lexical diversity of the dataset is quite moderate with distinct-1 being 0.091, distinct-2 0.375 and self-BLEU-4 0.441 (calculated on 300 random samples). Contact dermatitis stands out, with the lowest distinct-2 (0.282) and the highest self-BLEU-4 (0.746), which indicates templated phrasing, and with almost no texts naming the disease (1%, versus 46–87% for the other classes).

Table 4. Dataset statistics per class. Name mention is the share of texts that contain the class's own disease name. Distinct-2 and self-BLEU-4 (lower is more diverse) are computed within each class (self-BLEU-4 on up to 90 sampled texts).

Class	Texts	Name mention (%)	Distinct-2	Self-BLEU-4
Acne	99	87	0.484	0.452
Contact dermatitis	98	1	0.282	0.746
Eczema	90	74	0.512	0.421
Folliculitis	95	46	0.562	0.368
Psoriasis	100	61	0.536	0.441
Scabies	89	85	0.572	0.357
Vitiligo	82	59	0.520	0.395
All classes	653	58.7	0.375	0.441

5.2 Overall classification performance

Table 5 and Figure 2 summarise the results. In 5×5 cross-validation, Random Forest (RF) reaches 94.55 ± 0.55% accuracy and Multinomial Naïve Bayes (MNB) 91.58 ± 0.45%. On the pooled out-of-fold predictions of one cross-validation run, RF classifies 618 of 653 texts correctly (94.64%; Wilson interval 92.6–96.1%) and MNB 596 of 653 (91.27%; 88.9–93.2%). RF is higher on every aggregate metric. A paired exact McNemar test on the pooled predictions finds 37 texts that only RF classifies correctly and 15 that only MNB classifies correctly ($p = 0.003$), so the RF advantage is statistically supported.

Table 5. Overall performance. Cross-validation (CV) rows use 5 repeats of stratified 5-fold CV; pooled rows use the out-of-fold predictions of repeat 1 (n = 653). Single-split rows use a random 70/30 split (457 training and 196 test texts).

Metric	Random Forest	Multinomial NB
5×5 CV accuracy (%), mean ± sd	94.55 ± 0.55	91.58 ± 0.45
5×5 CV macro-F1, mean ± sd	0.944 ± 0.006	0.915 ± 0.005
Pooled accuracy, correct / total	618 / 653 (94.64%)	596 / 653 (91.27%)
Pooled 95% Wilson interval (%)	[92.6, 96.1]	[88.9, 93.2]
Pooled macro precision / recall / F1	0.945 / 0.945 / 0.945	0.913 / 0.912 / 0.912
Pooled weighted F1	0.946	0.912
Pooled Cohen's κ / MCC	0.937 / 0.937	0.898 / 0.898
Pooled macro one-vs-rest AUC	0.997	0.993
Paired McNemar test (pooled)	RF only correct: 37; MNB only: 15; p = 0.003	
20 random 70/30 splits, accuracy (%)	94.31 ± 1.47 (range 90.8–96.9)	90.54 ± 1.65 (range 87.8–93.4)
Single split, random_state = 42 (n = 196)	92.86% (182/196) [88.4, 95.7]	89.80% (176/196) [84.8, 93.3]
McNemar test on the single split	RF only correct: 14; MNB only: 8; p = 0.29	

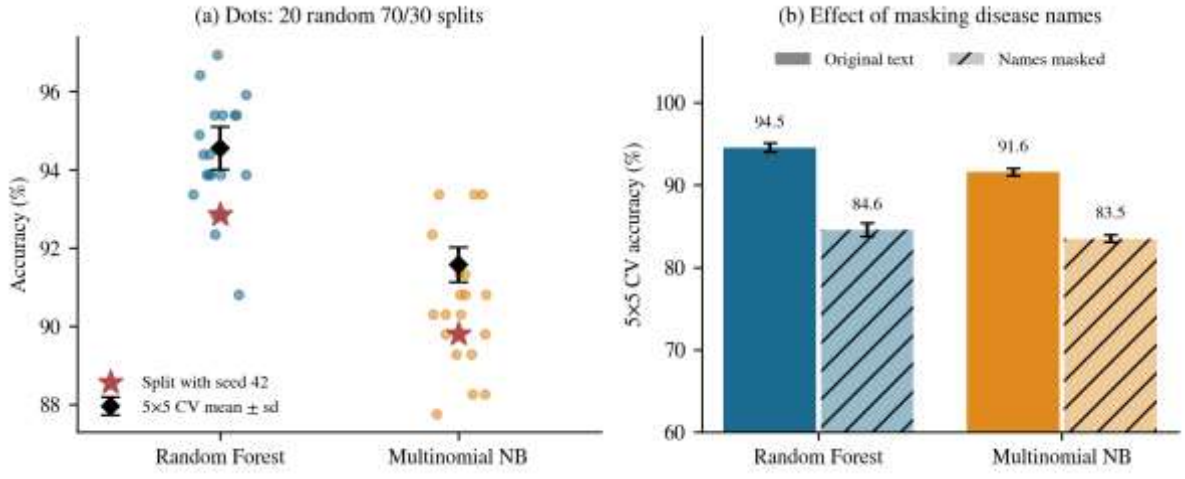


Fig. 2. (a) Accuracy of 20 random 70/30 splits (dots), the 5×5 CV mean ± sd (diamonds) and the split with seed 42 (stars). (b) 5×5 CV accuracy with the original text and with disease names masked (Section 5.5).

The single 196-text split with seed 42 gives 92.86% (RF) and 89.80% (MNB). Both values lie within the split-to-split variation (RF 90.8–96.9%, MNB 87.8–93.4% over 20 random splits), and on that split the McNemar test cannot separate the two models

(p = 0.29). A single split of this size should therefore not be used to rank classifiers or to compare prompting settings; the cross-validated estimates are more reliable.

5.3 Per-class performance

Table 6 and Figures 3–5 give per-class results from the pooled cross-validation predictions. RF reaches F1 ≥ 0.94 for five of the seven classes; its weakest classes are vitiligo (F1 0.896, recall 0.890) and eczema (F1 0.902, precision 0.883). MNB is weakest on eczema (F1 0.828, recall 0.800) and has low acne precision (0.853) because eight eczema texts are labelled acne. One-vs-rest AUC is at least 0.979 for every class under both models (macro 0.997 for RF and 0.993 for MNB). Contact dermatitis is the easiest class (RF F1 0.980), which is consistent with its templated, low-diversity phrasing (Section 5.1).

Table 6. Per-class precision (P), recall (R), F1 and one-vs-rest AUC from pooled 5-fold cross-validation predictions (n = 653).

Class	n	RF P	RF R	RF F1	RF AUC	NB P	NB R	NB F1	NB AUC
Acne	99	0.979	0.939	0.959	0.999	0.853	0.939	0.894	0.993
Contact dermatitis	98	0.970	0.990	0.980	1.000	0.933	0.990	0.960	1.000
Eczema	90	0.883	0.922	0.902	0.992	0.857	0.800	0.828	0.979
Folliculitis	95	0.948	0.968	0.958	0.995	0.946	0.916	0.930	0.991
Psoriasis	100	0.980	0.970	0.975	0.998	0.956	0.870	0.911	0.996
Scabies	89	0.954	0.933	0.943	0.997	0.946	0.978	0.961	0.998
Vitiligo	82	0.901	0.890	0.896	0.996	0.901	0.890	0.896	0.997

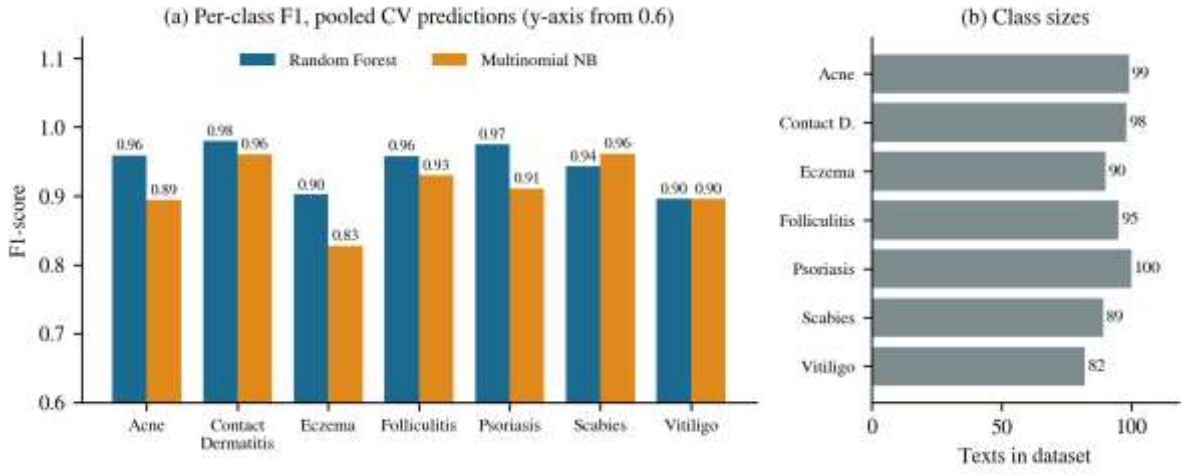


Fig. 3. (a) Per-class F1 from pooled cross-validation predictions (y-axis starts at 0.6). (b) Number of texts per class.

Natural Resources for Human Health
eISSN: 2583-1194

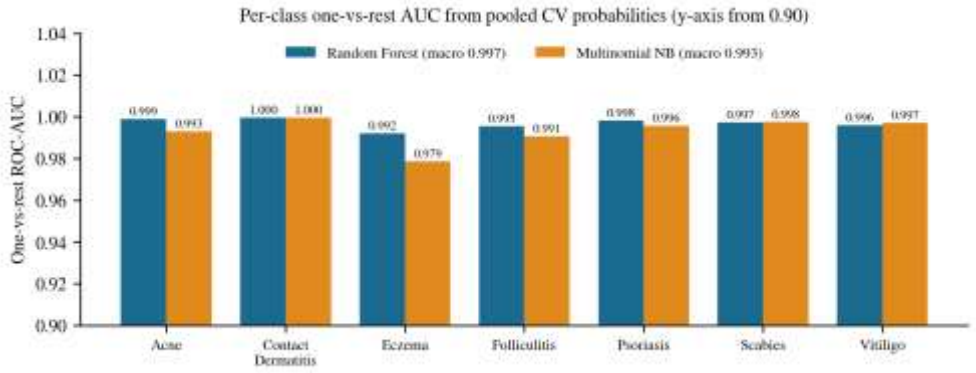


Fig. 4. Per-class one-vs-rest AUC from pooled cross-validation probabilities (y-axis starts at 0.90).

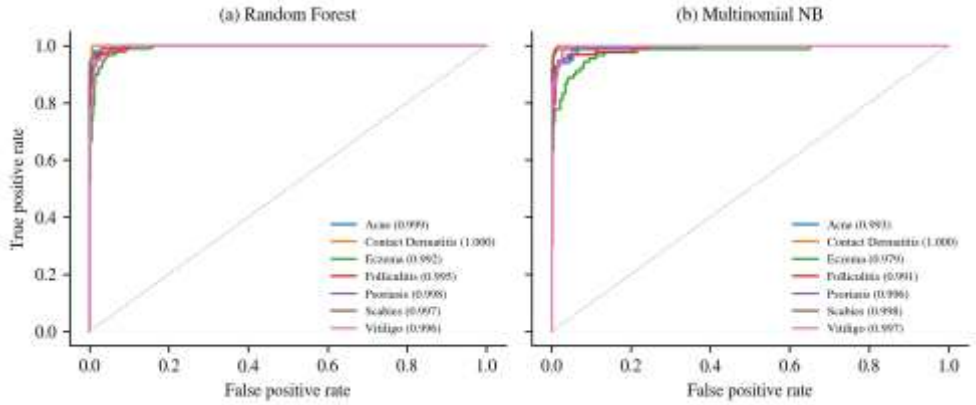


Fig. 5. One-vs-rest ROC curves from pooled cross-validation probabilities for (a) Random Forest and (b) Multinomial Naïve Bayes. Legend values are per-class AUC.

5.4 Confusion and error analysis

Figure 6 shows the pooled confusion matrices and Figure 7 the most frequent confusion pairs. RF makes 35 errors. The largest confusion is between vitiligo and eczema (vitiligo → eczema 9, eczema → vitiligo 5), which accounts for 14 of the 35 errors (40%); eczema is involved in 18 of them. MNB makes 57 errors, and eczema is involved in 30 of them: 18 eczema texts are missed (eight labelled acne, five vitiligo, two contact dermatitis, two scabies, one psoriasis) and 12 other texts are labelled eczema. Other frequent MNB confusions are psoriasis → eczema (6), vitiligo → eczema (4) and acne → folliculitis (3).

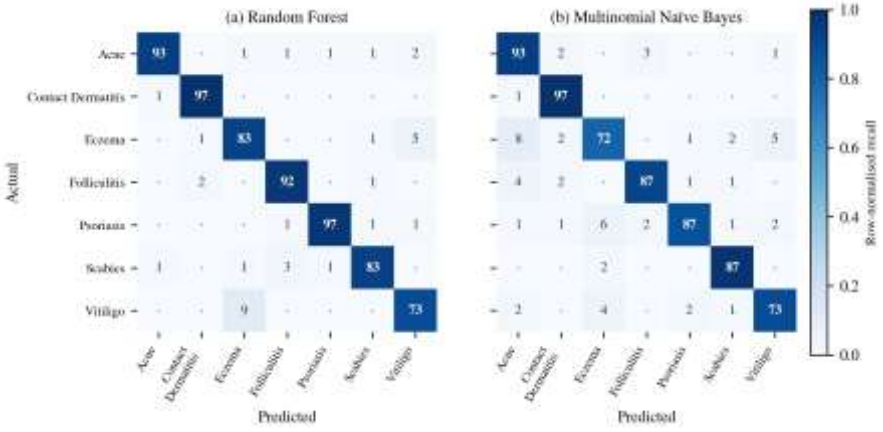


Fig. 6. Confusion matrices from pooled 5-fold cross-validation predictions (n = 653) for (a) Random Forest and (b) Multinomial Naïve Bayes. Cell values are counts; shading shows row-normalised recall.

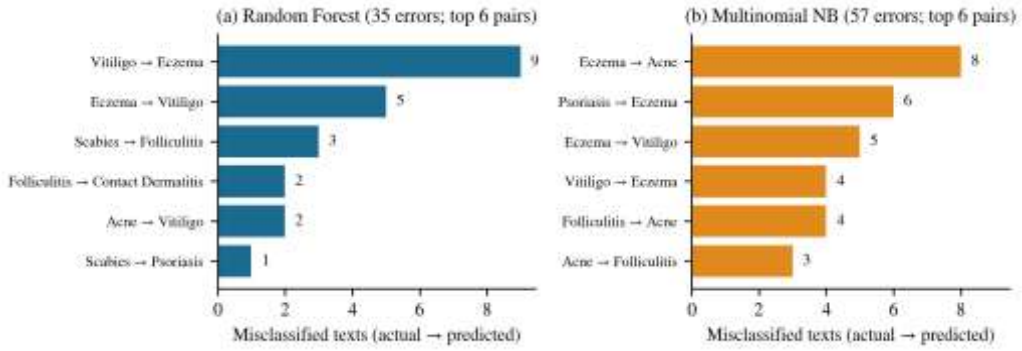


Fig. 7. Most frequent confusion pairs (actual → predicted) from pooled cross-validation predictions.

Table 7 lists misclassified RF predictions. The vitiligo and eczema errors are mostly short third-person clinical sentences that contain no disease name and few discriminative terms (for example "Depigmented areas are noticeable around the eyes and mouth."). This suggests that the model leans on disease names and class-specific vocabulary, and that generic clinical sentences are harder to classify. This is an observation from examples, not a tested explanation. Per-class rankings also differ on the seed-42 split (there, for instance, folliculitis has the lowest precision for MNB), another sign that per-class conclusions from one 196-text split are unstable.

Table 7. Examples of misclassified texts (Random Forest, pooled cross-validation predictions).

True label	Predicted	Text
Vitiligo	Eczema	Depigmented areas are noticeable around the eyes and mouth.
Vitiligo	Eczema	The skin patches have no itching, redness, or scaling.
Eczema	Vitiligo	There are cracks and bleeding in the skin folds of the fingers due to dryness.
Eczema	Vitiligo	Red, scaly patches are visible on the patient's elbows and knees.
Scabies	Folliculitis	I have tiny red bumps on my waistline and underarms. They itch so badly. What should I do?

5.5 Keyword-leakage ablation

Because 58.7% of the texts name their own disease (Table 4), we repeated the cross-validation after replacing every disease name (acne, eczema, atopic dermatitis, contact dermatitis, dermatitis, folliculitis, psoriasis, scabies, vitiligo) with a placeholder token. Masking lowers 5×5 CV accuracy from 94.55% to 84.56% for RF (−10.0 points) and from 91.58% to 83.49% for MNB (−8.1 points) (Table 8). The RF–MNB gap shrinks to 1.1 points and is no longer significant on pooled predictions (McNemar: RF only 54, MNB only 35, $p = 0.056$). The F1 drop is largest for classes whose texts most often name the disease: scabies (−0.157), folliculitis (−0.136), eczema (−0.106) and psoriasis (−0.100), and absent for contact dermatitis (+0.005), where names are almost never used (Figure 8).

Table 8. Effect of masking disease names on 5×5 cross-validation performance (mean ± sd).

Model	Accuracy, original (%)	Accuracy, masked (%)	Change (points)	Macro-F1 original / masked
Random Forest	94.55 ± 0.55	84.56 ± 0.84	−10.0	0.944 / 0.844

Model	Accuracy, original (%)	Accuracy, masked (%)	Change (points)	Macro-F1 original / masked
Multinomial NB	91.58 ± 0.45	83.49 ± 0.46	−8.1	0.915 / 0.833

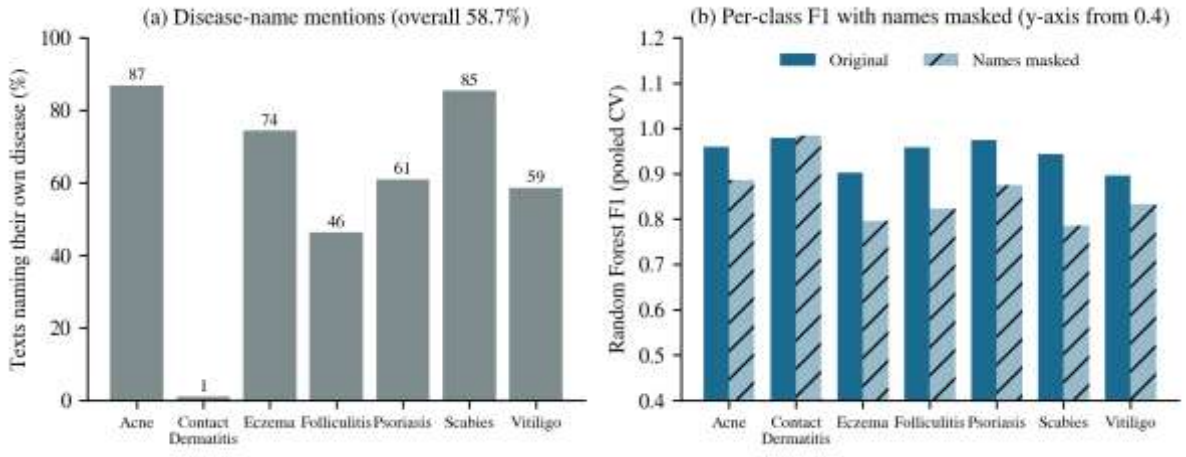


Fig. 8. (a) Share of texts that name their own disease, per class. (b) Random Forest per-class F1 with the original text and with disease names masked (pooled cross-validation predictions; y-axis starts at 0.4).

Table 9 shows the terms that most strongly favour each class in a Multinomial NB model fitted on all texts (terms occurring at least eight times in the class, ranked by log-odds against the other classes). The disease name is the top term for six of the seven classes. Several other top terms look like generation artefacts rather than clinical vocabulary, for example "his" and "woman" for folliculitis and "ups" (from "flare-ups") for eczema, which again points to templated narratives.

Table 9. Top six terms per class, ranked by Multinomial NB log-odds against the other classes.

Class	Top terms
Acne	acne, breakouts, scarring, whiteheads, pimples, include
Contact dermatitis	reaction, allergic, fabric, cleaning, new, gloves
Eczema	eczema, unbearable, ups, flare, gets, sure
Folliculitis	folliculitis, look, tender, his, bumps, woman
Psoriasis	psoriasis, thick, find, bleed, better, feel
Scabies	scabies, scratching, fingers, crusted, intense, confirmed
Vitiligo	vitiligo, white, forming, spreading, spread, areas

The masked results are a more conservative estimate of how well the classifiers use symptom descriptions, but they are still obtained on synthetic text and may still rely on templated phrasing. We therefore recommend quoting both the original and masked figures.

5.6 Learning curve

Figure 9 depicts held-out accuracy for increasing training samples for each class, starting from 10 up to 60 (5-fold cross validation). RF's accuracy increases from 87.8% to 94.2%, while that of MNB is increased from 77.2% to 90.7%. It is not yet clear whether the graphs have flattened out; hence, more synthetic data can be of use, especially MNB.

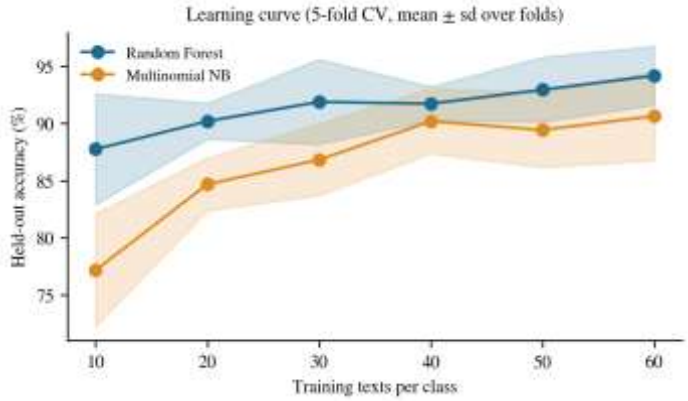


Fig. 9. Learning curve: accuracy versus number of training texts per class (5-fold cross-validation; mean ± sd over folds).

5.7 Expert evaluation

Table 10 and Figure 10 present the mean dermatologist ratings. The average across the four criteria is 4.53/5, with the highest score for medical coherence (4.7) and the lowest for realism (4.3).

Table 10. Mean ratings of the synthetic texts by three dermatologists (scale 1–5).

Criterion	Mean score
Medical coherence	4.7
Symptom accuracy	4.6
Diversity	4.5
Realism	4.3
Mean across criteria	4.53

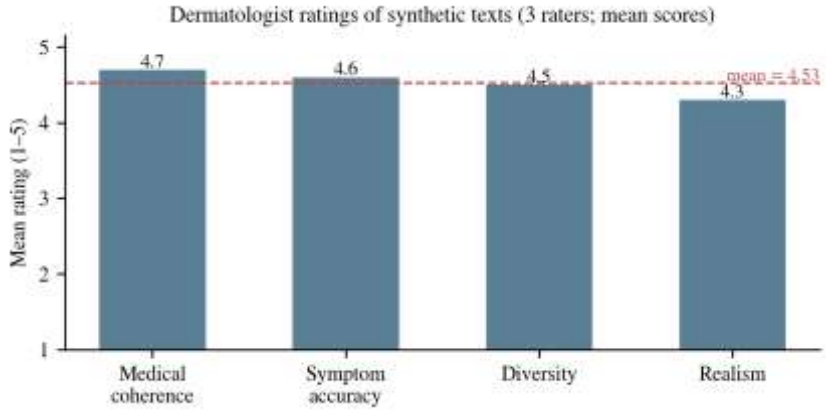


Fig. 10. Mean dermatologist ratings per criterion (3 raters; only mean scores are available, so no dispersion is shown).

The qualitative feedback supported the realism score since it revealed that the text is quite well-structured compared to the description of actual patients, it may not contain phrases used by actual patients, it is good at capturing common diseases but

does not cover rare disease types. These comments are consistent with the quantitative feedback in the sense that the text contains short, templated and keyword-filled sentences. Since the feedback is based on means of only three raters, it lacks any measure of dispersion or agreement.

6. DISCUSSION AND LIMITATIONS

It only took a set of 67 examples and a well-crafted prompt for GPT-4 to generate 653 texts that allow classifiers with 91-95% accuracy and dermatologists believe are coherent and clinically sound. The analysis also shows that this accuracy overstates what the classifiers learn about symptoms. The main limitations are as follows.

- **Synthetic-only evaluation.** Cross-validation has been done for each classification outcome using artificial text. The transfer to actual patient text has never been evaluated, so do not interpret the results as real-world diagnostic outcomes.
- **Keyword leakage.** 58.7% of the texts mention their disease, and removing the disease name reduces the accuracy by 8-10 points. The future generation must either tell the model not to mention the disease or use a combination of mentioned and non-mentioned texts.
- **Templating and short length.** The texts have an average of 21.7 words, the condition contact dermatitis is very repetitive (self-BLEU-4 0.746), and the most frequent terms are generation artifacts. Texts that are longer and varied are required. The folds in cross-validation can use the same generation examples; this can inflate the scores slightly.
- **Single-split instability.** A single 196-text split cannot separate the classifiers ($p = 0.29$) and varies by several points between splits. Cross-validated and paired analyses should be used.
- **Exemplar design and per-k analysis.** Exemplars were not class-conditioned (Section 4.3), and the dataset file does not record which k produced each text, so the effect of k cannot be measured. Adding a k column would allow a per-k ablation.
- **Seed size and provenance.** The training is based on just 67 publicly available texts. Rare subtype coverage, demographic variation, and non-English text is not known. In addition, the experts mentioned that rare subtypes were underrepresented.
- **Privacy.** As the seed texts were publicly available, there were no patient-level private information used, and therefore the privacy advantage could not be tested. However, if the DermaSynth tool is utilized on privately-owned clinical texts, the risk of memorization and de-identification should be evaluated, such as by using nearest neighbor similarity.
- **Generator reproducibility.** GPT-4 is a proprietary, versioned and stochastic model; results may not reproduce exactly. Releasing the generated dataset and prompts mitigates this.
- **Scope of clinical claims.** The dataset contains patient-style narratives and clinical descriptions, not validated diagnoses. Text classification accuracy is not diagnostic performance, and the work is not a clinical decision tool.
- **Expert study.** Three raters, mean scores only, and no reported agreement statistic.

7. CONCLUSION AND FUTURE WORK

DermaSynth generates 653 synthetic labeled texts from 67 dermatology texts through GPT-4 with context, diversity, and few-shot prompting. With 5×5 cross-validation, Random Forest achieves 94.55% accuracy while Multinomial Naïve Bayes achieves 91.58%. This accuracy gap is statistically significant (McNemar's $p=0.003$). However, 58.7% of the texts have named diseases, which when masked reduces the accuracy to 84.6% and 83.5% respectively, and the texts are also relatively short with very template-like texts for contact dermatitis.Dermatologists rated the data 4.53/5 on average. The study supports the feasibility of prompt-based synthetic dermatology text, while showing that keyword cues and the absence of a real-text test limit the strength of the conclusions.

Future work includes: (i) train-on-synthetic/test-on-real evaluation on a larger real corpus; (ii) generation that avoids naming the disease, with longer and more varied texts; (iii) a per-k ablation with the k used for each text recorded, and class-conditioned exemplar selection; (iv) stronger classifiers, including fine-tuned encoder models such as BioBERT or ClinicalBERT used as classifiers and embedding extractors; (v) multi-generator comparisons with open-weight LLMs; (vi) prompts that add differentiating symptoms for confusable pairs such as eczema and vitiligo; and (vii) multilingual extension.

Declarations

Ethics. The expert evaluation rated synthetic text and did not involve patients or patient records.

REFERENCES

[1] [1] R. J. Hay, N. E. Johns, H. C. Williams, I. W. Bolliger, R. P. Dellavalle, D. J. Margolis, R. Marks, L. Naldi, M. A. Weinstock, S. K. Wulf, C. Michaud, C. J. L. Murray, and M. Naghavi, "The global burden of skin disease in 2010: An analysis of the prevalence and impact of skin conditions," J. Invest. Dermatol., vol. 134, no. 6, pp. 1527–1534, 2014.

[2] F. J. Dalgard, U. Gieler, L. Tomas-Aragones, B. Lien, F. Poot, G. B. E. Jemec, et al., "The psychological burden of skin diseases: A cross-sectional multicenter study among dermatological out-patients in 13 European countries," J. Invest. Dermatol., vol. 135, no. 4, pp. 984–991, 2015.

[3] Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," Nature, vol. 542, no. 7639, pp. 115–118, 2017, doi:10.1038/nature21056.

[4] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Sci. Data, vol. 5, art. 180161, 2018, doi:10.1038/sdata.2018.161.

[5] "Skin disease text classification," Kaggle dataset (user rafsunahmad). [Online]. Available: https://www.kaggle.com/datasets/rafsunahmad/skin-disease-text-classification

[6] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 1877–1901, arXiv:2005.14165.

[7] OpenAI, "GPT-4 technical report," arXiv:2303.08774, 2023.

[8] C. Peng et al., "A study of generative large language model for medical research and healthcare," npj Digit. Med., vol. 6, art. 210, 2023.

[9] S. Kweon et al., "Publicly shareable clinical large language model built on synthetic clinical notes," in Findings of the Association for Computational Linguistics: ACL 2024, 2024, arXiv:2309.00237.

[10] G. Kumichev, P. Blinov, Y. Kuzkina, V. Goncharov, G. Zubkova, N. Zenovkin, A. Goncharov, and A. Savchenko, "MedSyn: LLM-based synthetic medical text generation framework," in Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2024), 2024, doi:10.1007/978-3-031-70381-2_14, arXiv:2408.02056.

[11] Z. Li, H. Zhu, Z. Lu, and M. Yin, "Synthetic data generation with large language models for text classification: Potential and limitations," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2023, arXiv:2310.07849.

[12] Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," Nature, vol. 631, pp. 755–759, 2024.

[13] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak, "Automated encoding of clinical documents based on natural language processing," J. Am. Med. Inform. Assoc., vol. 11, no. 5, pp. 392–402, 2004.

[14] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, and D. Sontag, "Large language models are few-shot clinical information extractors," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 1998–2022.

[15] Y. Labrak, M. Rouvier, and R. Dufour, "A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks," in Proc. LREC-COLING, 2024, arXiv:2307.12114.

[16] S. Belkadi, L. Ren, N. Micheletti, L. Han, and G. Nenadic, "Generating synthetic free-text medical records with low re-identification risk using masked language modeling," in Proc. NAACL 2025 Student Research Workshop, 2025, arXiv:2409.09831.

[17] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-Instruct: Aligning language models with self-generated instructions," in Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL), 2023, arXiv:2212.10560.

[18] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification," Neurocomputing, vol. 321, pp. 321–331, 2018.

[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[20] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[21] McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in AAAI-98 Workshop on Learning for Text Categorization, 1998, pp. 41–48.

[22] E. B. Wilson, "Probable inference, the law of succession, and statistical inference," J. Am. Stat. Assoc., vol. 22, no. 158, pp. 209–212, 1927.
[23] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947.
[24] Cohen, "A coefficient of agreement for nominal scales," Educ. Psychol. Meas., vol. 20, no. 1, pp. 37–46, 1960.
[25] Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," Comput. Biol. Chem., vol. 28, no. 5–6, pp. 367–374, 2004.