Augmentation of ATR-FTIR Spectra in Biomedical Fluids: A Baseline-Centered, Geometry-First Evaluation with Stability and Synthetic Data Quality Control

Margarita Balandina
Moscow Institute of Physics and Technology
(National Research University)
Moscow, Russia

Tatiana Litvinova
Belgorod State University,
Belgorod, Russia
Voronezh State Pedagogical University
Voronezh, Russia

Abstract—Augmenting infrared spectra is increasingly used to mitigate data scarcity, yet on small clinical datasets it may also create misleading improvements if evaluated without a strict baseline and synthetic quality control (QC). We treat augmentation as a methodological experiment and compare a baseline (training on original data only) against train-only spectrum-level augmentations (Gaussian noise, spectral shift and Mixup) in three biomedical use cases: 1) gingival crevicular fluid (GCF, n=18) with multi-factor clinical annotation, 2) saliva COVID-19 screening (patient-level protocol), and 3) saliva diabetes screening (holdout protocol). Beyond supervised metrics (Recall/F1 for pathology, PR-AUC, specificity), we evaluate probability calibration (Brier score and expected calibration error, ECE) and, crucially, the geometry of the feature space via PCA. For the small-n GCF dataset, we quantify factor–coordinate association strength using ANOVA- and correlation-based measures and demonstrate that augmentation can rotate the PCA basis; therefore, comparisons must rely on best-PC or top-k criteria rather than fixed PC1/PC2. The key finding on the small-n GCF dataset is geometric rather than clinical or biological. Within a predefined narrow spectral window frequently used in prior GCF studies, classic augmentation redistributes variance across higher-order PCA components, leading to an increase in the maximum factor–PC association metric ($\Delta\max R^2 \approx +0.121$ on average), which reflects variance redistribution across rotated PCA axes rather than improved separability. This effect is negligible on broader spectral ranges, highlighting the localization and instability of augmentation effects under small-n conditions. Overall, augmentation should be treated as an experiment on data geometry and model stability, and interpreted only together with strict baseline comparison, synthetic QC, and sanity checks—especially in small clinical datasets.

Keywords—ATR-FTIR; gingival crevicular fluid; saliva; data augmentation; PCA; calibration; quality control.

I. INTRODUCTION

Augmentation of Attenuated Total Reflection Fourier Transform Infrared Spectra (ATR-FTIR) spectroscopy provides a rapid, reagent-free molecular fingerprint of biological fluids. However, many biomedical FTIR studies remain small and heterogeneous, especially for gingival crevicular fluid (GCF), where only microliter volumes are available and cohorts are typically limited. In such settings, supervised classifiers are high-variance, and naive reports of "improved metrics after augmentation" are not convincing without:

- a strict baseline
- train-only application to prevent leakage
- synthetic QC and structure-preserving sanity checks

Despite the growing use of augmentation in spectral machine learning, its reported benefits remain difficult to compare across biomedical FTIR studies because datasets differ not only in size, but also in annotation structure, repeated measurements, and leakage risk. This is especially important for clinical spectra, where apparent metric gains may reflect either a genuine regularization effect or an evaluation artifact. Therefore, the goal of this study is not simply to ask whether augmentation can improve a metric in a single task, but to determine under which data conditions and evaluation protocols such gains remain trustworthy.

Recent GCF studies have used exploratory multivariate analysis (PCA/HCPC) to relate spectral variability to clinical factors and highlighted the Amide III region as an informative window [1, 2]. In this work, we deliberately focus on methodological aspects of spectral augmentation and explicitly avoid biological or clinical interpretation of principal components or spectral features. Previously published studies have already explored clinical associations in GCF spectra using exploratory multivariate analysis; here, PCA is used strictly as a tool to monitor data geometry and stability under augmentation.

Our contributions are:

- a unified baseline-to-augmentation protocol across small-n GCF and larger saliva datasets;
- a geometry-first evaluation using PCA and factor–coordinate association measures with a best-PC/top-k comparison rule to account for augmentation-induced basis rotation
- unsupervised sanity checks (clustering and resampling stability) and synthetic QC metrics
- a reproducible bash/python pipeline with fixed seeds, configuration logging and automatic aggregation of reports

II. Materials and Methods

A. Datasets.

The datasets were selected purposively to cover complementary methodological regimes relevant to augmentation analysis rather than to represent a single disease domain. The selection criteria were: biomedical relevance of ATR-FTIR analysis, availability of raw spectra and labels or metadata sufficient for reproducible preprocessing and task definition, feasibility of leakage-safe evaluation, either through patient identifiers or a clearly defined sample-level protocol, and contrast in dataset scale. Under these criteria, GCF was included as an extreme small-n clinically annotated dataset, COVID saliva as a patient-level repeated-spectra dataset, and diabetes saliva as a larger sample-level dataset used to test whether augmentation effects become more stable with increasing scale.

1) Gingival crevicular fluid (GCF, small-n): 18 spectra (p = 1781). Metadata: sex (Gender), age group (Age_factor), caries status (caries_factor), periodontal status (Parodont), anamnesis group (Anamnes_factor). We consider three binary classification tasks derived from the clinical labels:

- periodontally healthy vs periodontal pathology
- healthy anamnesis vs anamnesis with pathology/conditions
- Overall healthy vs any pathology

2) Saliva (COVID-19): 183 spectra from 61 subjects; evaluation uses patient-level Monte-Carlo double cross-validation (MCDCV) to avoid leakage across spectra from the same subject.

3) Saliva (diabetes): 1040 spectra; holdout-CV protocol. Patient IDs are not available, therefore results are interpreted at sample level; stratification controls are applied to reduce confounding.

B. Baseline vs augmentation. Baseline uses the same preprocessing, splitting protocol and models but without synthetic expansion. Classic augmentation applies Gaussian noise, spectral shift and Mixup [4] strictly inside training folds (train-only). For calibration we use Platt scaling [5]; performance is reported for the pathology class using Recall and F1, along with PR-AUC and specificity; probability quality is assessed by Brier score [7] and ECE [6].

C. Geometry and structure.

We run PCA on the real samples and quantify factor–coordinate association measures between each clinical factor and PCA coordinates: η2 from one-way ANOVA for categorical factors and R^2 (squared correlation) for continuous factors. PCA is used here strictly as a tool to assess data geometry and stability under augmentation, not for biological or clinical interpretation.

Because augmentation can rotate the PCA basis, we compare baseline vs augmented representations using best-PC per factor (maximum association over PCs) or top-k PCs, rather than fixed PC1/PC2. This redistribution is illustrated in Fig. 1 for a representative factor in the Amide III window.

D. Unsupervised sanity checks and QC. Clustering is performed on PCA scores using KMeans, Gaussian mixture models and agglomerative clustering (K=2…5). We report internal indices (silhouette, Davies–Bouldin) and resampling stability measured by adjusted Rand index (ARI). Synthetic QC includes a real-vs-synthetic domain classifier (AUC_dev=|AUC−0.5|, lower is better, 0 means indistinguishable), kNN domain overlap, and distribution-shift measures based on within-real (rr) vs real-to-synthetic (rs) distance statistics and Wasserstein distance between rr and rs distributions [3].

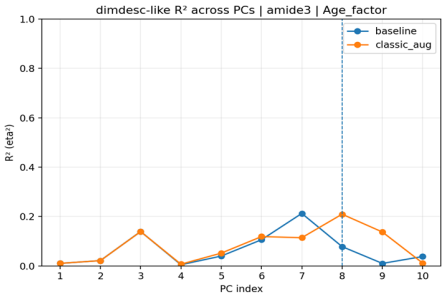

Fig. 1. Amide III: Age Factor (baseline vs classic augmentation).

E. Reproducibility. Experiments are executed via a unified bash/python pipeline that fixes random seeds, stores run configurations, keeps train/test splits group-aware where IDs are available, and exports per-run and aggregated CSV reports (mean±std over repeated runs). This enables direct baseline vs augmentation comparisons under identical preprocessing and evaluation protocols.

III. Results

A. Saliva datasets: supervised performance and calibration

Table I summarizes augmentation effects (Aug−Base, augmentation minus baseline) averaged over models. For COVID-19, classic augmentation yields small but consistent gains in Recall and PR-AUC, with an expected specificity trade-off. For diabetes, strong augmentation produces a larger and more stable improvement, notably for SVM-RBF. All effects are reported as Aug−Base under identical preprocessing and evaluation protocols.

TABLE I. AUGMENTATION EFFECTS

Dataset	Model/avg	ΔPR-AUC	ΔRecall	ΔF1	ΔSpec	ΔBrier / ΔECE
COVID	LDA	+0.006	+0.004	-0.001	-0.013	-0.004 / -0.006
COVID	LOGREG	+0.010	+0.053	+0.026	-0.027	-0.009 / -0.001
COVID	PLSDA	+0.001	+0.003	-0.001	-0.010	-0.001 / +0.002
COVID	SVM_LIN	+0.011	+0.021	+0.006	-0.026	-0.005 / -0.001
COVID	SVM-RBF	+0.005	+0.037	+0.025	-0.010	-0.004 / +0.001
COVID	Mean	+0.007	+0.024	+0.011	-0.017	-0.004 / -0.001
Diabetes*	LDA	-0.003	+0.006	+0.006	+0.006	-0.005 / -0.001
Diabetes*	LOGREG	+0.009	+0.034	+0.013	-0.013	-0.013 / -0.004
Diabetes*	PLSDA	+0.001	-0.009	-0.001	+0.014	-0.001 / -0.002
Diabetes*	SVM-RBF	+0.048	+0.083	+0.059	+0.037	-0.049 / -0.030
Diabetes*	Mean	+0.014	+0.028	+0.020	+0.011	-0.017 / -0.009

[1] *No patient IDs; sample-level interpretation

B. *Small-n GCF: augmentation effects on data geometry*

The core result on the GCF dataset (n = 18) is geometric rather than clinical. We evaluate how augmentation alters the structure of PCA representations and factor–coordinate association measures. The Amide III window (1185–1330 cm^{-1}) is treated as an a priori predefined spectral interval frequently used in prior GCF studies [1,2] and is used here strictly for methodological comparison. No biological or clinical interpretation of PCA components is performed.

Using best-PC aggregation across repeated runs (n = 5 seeds), classic augmentation increases the average maximum factor–PC association metric within the Amide III window by $\Delta\max R^2 \approx +0.121$ (median +0.114), whereas the effect is close to zero on broader spectral ranges (paper_full: +0.002; paper_low: +0.004), see Fig. 2. This localization indicates that augmentation effects on small-n GCF primarily manifest as redistribution of variance across higher-order PCA components rather than as a global strengthening of structure.

Negative ΔR^2 values on individual components are expected and reflect PCA basis rotation under augmentation rather than loss of structure. Factor-level values are reported in Table II.

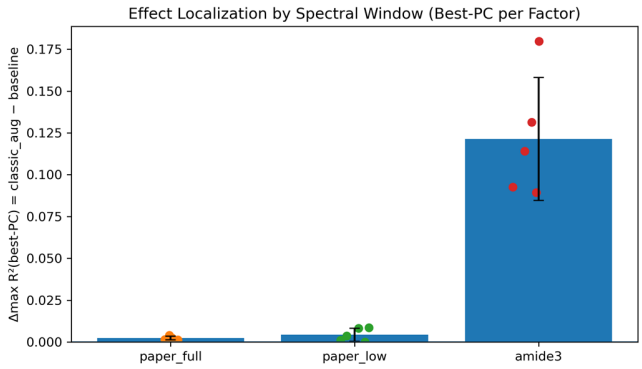

Fig. 2. Effect Localization by Spectral Window (Best-PC per Factor)

TABLE II. FACTOR-PC ASSOCIATION METRICS IN THE AMIDE III WINDOW (BEST-PC PER FACTOR

Factor	Best PC	R2 base	R2 aug	ΔR2	ΔR2 std
Anamnes_factor	PC9	0.013	0.193	+0.180	0.039
Age_factor	PC8	0.078	0.210	+0.131	0.057
caries_factor	PC10	0.115	0.229	+0.114	0.035
Gender	PC10	0.010	0.103	+0.093	0.025
Parodont	PC9	0.093	0.182	+0.089	0.100

C. *Unsupervised sanity checks and synthetic QC on GCF*

On the broad-range profile (paper_full), the intrinsic cluster structure is simple (best k=2 by silhouette). Fitting PCA on (real + classic_aug) and projecting real samples preserves the structure and can improve resampling stability for some algorithms without altering the intrinsic cluster structure (Table III). QC metrics quantify the domain shift introduced by classic augmentation.

As a QC-only contrast, we additionally report VAE-based synthetic samples to illustrate risks of generative synthesis under small-n**;** VAE shows substantially larger global mismatch than classic perturbations. QC results are summarized in Table IV.

TABLE III. CLUSTERING QUALITY AND RESAMPLING STABILITY ON PCA SCORES (PAPER_FULL, K=2; MEAN OVER 80 RESAMPLES)

Algorithm	Silhouette (base→aug)	DB (base→aug)	ARI (base→aug)
AGGLO	0.589→0.604	0.337→0.297	0.696→0.758
KMEANS	0.582→0.585	0.356→0.352	0.680→0.701
GMM	0.597→0.591	0.299→0.296	0.701→0.680

TABLE IV. SYNTHETIC QC SUMMARY ON GCF (CLASSIC AUGMENTATION; VAE SHOWN FOR QC-ONLY CONTRAST)

QC metric	Classic aug	VAE (QC only)	Note
AUC_dev (real vs synth)	0.091	0.226	Domain distinguishability; lower is better.
kNN overlap	0.299	0.316	Local domain mixing.
Wasserstein(rr,rs) median	1.415	5.897	Global shift (higher is worse).
Effective PCA dims (mean)		3.324	Low dimension indicates under-dispersion.

Supervised results on GCF are intentionally reported compactly as a behavioral check. Across tasks, augmentation effects are not universal: some settings improve pathology Recall/F1, but cases of false optimism appear (Recall increases while specificity and calibration deteriorate), which reinforces the need for baseline comparison, QC and stability analysis before any clinical claims.

IV. Discussion

Our experiments suggest that augmentation effects are size-dependent. On larger saliva datasets, train-only augmentation acts mostly as a regularizer: it improves Recall/F1/PR-AUC modestly for COVID-19 and more strongly for diabetes, while typically improving calibration (Brier/ECE). In contrast, on small-n GCF, classic augmentation leads to increased factor–PC association metrics within the predefined Amide III window due to redistribution of variance across PCA components, rather than robust factor strengthening.

Methodologically, the best-PC/top-k rule is essential, because augmentation changes the PCA basis and can shift which component carries a given factor. Finally, QC separates perturbation-based augmentation from more risky generative synthesis: even when synthetic samples look locally close, global shift and low intrinsic dimension may indicate under-dispersion.

Practical takeaway: 1) report augmentation effects only as Aug−Base under a strict train-only protocol; 2) for PCA-based analysis, compare best-PC or top-k representations rather than fixed PC1/PC2; 3) always include QC and resampling-based stability checks; 4) expect stronger and more stable gains on larger datasets, while small-n settings require cautious, task-specific interpretation.

V. Limitations and Future Work

Limitations include the extremely small GCF cohort (n=18) and heterogeneous clinical factors, which make supervised results high-variance and therefore exploratory. In such small-n settings, augmentation effects are strongly representation-dependent, and apparent improvements may reflect geometric redistribution rather than robust signal enhancement.

Future work will focus on identifying conditions under which augmentation effects become invariant across

representations and evaluation protocols. This includes testing domain-specific and constraint-based augmentations, external validation on independent cohorts, and the development of standardized QC benchmarks for generative synthesis on one-dimensional spectral data.

VI. Conclusion

We proposed a baseline-centered, geometry-first protocol to study FTIR augmentation under clinical small-n constraints. On the small-n GCF dataset, classic train-only augmentation consistently alters the geometry of PCA representations within the predefined Amide III window, leading to an increase in maximum factor–PC association metrics ($\Delta\max R^2 \approx +0.121$) that reflects redistribution of variance across PCA components rather than robust factor strengthening. In contrast, effects on broader spectral ranges are negligible. On larger saliva datasets, supervised performance and calibration gains are more stable.

Overall, augmentation should be treated as an experiment on data geometry and model stability, and interpreted only together with strict baseline comparison, synthetic QC, and sanity checks.

References

[1] P. Seredin, T. Litvinova, Y. Ippolitov, D. Goloshchapov, Y. Peshkov, V. Kashkarov, I. Ippolitov, and B. Chae, "A study of the association between primary oral pathologies (dental caries and periodontal diseases) using synchrotron molecular FTIR spectroscopy in view of the patient's personalized clinical picture (demographics and anamnesis)", *International journal of molecular sciences*, vol. 25, no. 12, 6395, 2024.

[2] P. Seredin, T. Litvinova, Y. Ippolitov, D. Goloshchapov, Y. Peshkov, B. Chae, R. O. Freitas, and F. C. B. Maia, "Multivariate spectroscopic analysis of protein secondary structures in gingival crevicular fluid: insights from FTIR Amide III band across oral disease stages", *International journal of molecular sciences*, vol. 26, no. 10, 4693, 2025.

[3] U. Blazhko, V. Shapaval, V. Kovalev, and A. Kohler, "Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra", *Chemometrics and Intelligent Laboratory Systems*, vol. 215, 104367, 2021.

[4] H. Zhang, M. Cisse, Y. Dauphin, and D. Lopez-Paz, "Mixup: beyond empirical risk minimization", *Proc. ICLR*, 2018.

[5] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods", in advances in large margin classifiers, 1999, pp. 61–74.

[6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks", Proc. ICML, 2017.

[7] G. W. Brier, "Verification of forecasts expressed in terms of probability", *Monthly weather review*, vol. 78, pp. 1–3, 1950.