Expanding AI-Based Optical Urinalysis with Synthetic Data: A Large-Scale 849-Patient Study

Igor Balashov [a,b], **Sofya Stroganova**[a], **Yuri Poimanov** [a], **Vahan Gevorgyan** [c], **Armen Charchyan**[c], **Andrey Grunin**[a], **Andrey Fedyanin**[a]

[a] *Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia*

[b] *Uray Technologies, Yerevan 0012, Armenia*

[c] *Izmirlian Medical Center, Yerevan 0014, Armenia*

1. Introduction

Optical transmission spectroscopy, combined with machine-learning approaches, offers a convenient and noninvasive method for detecting clinically significant markers in urine [1]. One of the major challenges in developing robust classifiers for urinalysis is the imbalance between relatively scarce "positive" (out-of-range) samples and a larger set of "negative" (within-range) samples [2]. Limited volumes of data and markedly skewed class distributions can hamper sensitivity to rare—but potentially critical—conditions.

To address these obstacles, we investigated synthetic data augmentation in the context of a 15-parameter clinical urinalysis study that included pH, specific gravity, protein, glucose, ketones, bilirubin, white blood cells, red blood cells (isomorphic), red blood cells (dysmorphic), mucus, casts, urobilinogen, bacteria, yeast, and epithelial cells. For illustrative purposes, we focus here on six parameters—namely pH, protein, ketones, bilirubin, urobilinogen, and bacteria—as they span a range of prevalence rates and imbalance levels. Two synthetic-data strategies (ratio-preserving expansion vs. balanced expansion) are tested to explore whether artificially generated spectra can improve classification performance [3].

2. Dataset and Methods

Data were collected from 849 urine specimens using a custom-built optical transmission spectrometer operating across ultraviolet, visible and near-infrared wavelengths. Each specimen was assigned a "positive" or "negative" label per clinically established cutoffs for the abovementioned parameters. Synthetic augmentation was adopted to boost the training set and potentially enhance model sensitivity to these minority classes. Specifically:

1. Ratio-preserving generation. Within each class (positive, negative), all pairwise averages of spectra were computed and appended, expanding both classes proportionally while maintaining the original ratio.

2. Balanced generation. Additional synthetic spectra were produced only (or primarily) for the minority class, aiming to approximate an even split of positive and negative samples.

All experiments employed a partial least squares (PLS) classifier, evaluated by 5-fold Stratified Group Cross-Validation (grouped by patient) [4]. Validation metrics—namely sensitivity, specificity, and their harmonic mean (F-score)—were computed on the held-out folds to gauge generalization performance.

Where appropriate, we acknowledge that confidence intervals for these metrics can provide additional insight into uncertainty; these intervals and the associated statistical considerations will be presented and discussed in detail during the oral session. Even when confidence intervals are accounted for, the improvement trend from synthetic data remains consistently positive.

3. Results

Table 1 summarizes, for six representative parameters, the counts before and after synthetic augmentation, along with the validation F-scores for: (a) no synthetic data (baseline); (b) ratio-preserving generation; and (c) balanced generation.

Analyzing these parameter-specific outcomes suggests a consistent pattern:
1) For parameters with very few positives (e.g., bilirubin, urobilinogen), ratio-preserving generation substantially boosts the F-score relative to both no generation and balanced generation.
2) For moderate positive-class sizes (e.g., protein or ketones), ratio-preserving again surpasses balanced expansion, though the gap can be narrower.
3) Even in cases with larger positive sets (e.g., bacteria), ratio-preserving yields the highest F-score overall.

This aligns with the notion that direct proportional expansion of both classes may preserve decision boundaries more effectively than artificially balancing extreme disparities—at least in the context of our PLS-based classification and the present cross-validation folds.

4. Discussion and Conclusion

The use of synthetic data demonstrated clear benefits across all tested parameters, with ratio-preserving generation consistently surpassing balanced expansion in validation F-scores. This advantage likely arises because class balancing results in a smaller overall increase in sample size compared to ratio preservation. These findings underscore the importance of exploring multiple augmentation strategies rather than solely focusing on balancing minority classes—particularly in multi-parameter urinalysis, where prevalence rates vary significantly.

The results suggest that more sophisticated synthetic data pipelines could enhance performance and systematically address challenges posed by imbalanced medical datasets. The observed improvements remain robust even when accounting for confidence intervals, indicating a reliable effect. A comprehensive discussion of confidence intervals and associated caveats will be provided during the oral session. Collectively, the results affirm that, within this dataset and PLS classification framework,

Expanding AI-Based Optical Urinalysis with Synthetic Data: A Large-Scale 849-Patient Study

Igor Balashov [a,b], **Sofya Stroganova**[a], **Yuri Poimanov** [a], **Vahan Gevorgyan** [c], **Armen Charchyan**[c], **Andrey Grunin**[a], **Andrey Fedyanin**[a]

[a] *Faculty of Physics, Lomonosov Moscow State University, Moscow 119991, Russia*

[b] *Uray Technologies, Yerevan 0012, Armenia*

[c] *Izmirlian Medical Center, Yerevan 0014, Armenia*

ratio-preserving augmentation yields the most substantial performance gains while offering opportunities for further refinement through advanced synthetic approaches.

Table 1: Validation F-Scores for Six Example Parameters under Different Synthetic Data Strategies

Parameter	Original (Pos / Neg)	Ratio-Preserving (Pos / Neg)	No Generation F-score	Ratio-Preserving F-score	Balanced F-score
pH	32 / 817	528 / 12720	0.681	0.810	0.785
Protein	196 / 653	19306 / 63903	0.825	0.841	0.832
Ketones	49 / 800	1225 / 19306	0.755	0.811	0.785
Bilirubin	10 / 839	55 / 3828	0.854	0.953	0.822
Urobilinogen	7 / 842	28 / 2556	0.885	0.968	0.943
Bacteria	269 / 580	36315 / 77815	0.815	0.828	0.825

References

[1] P. Sokołowski, K. Cierpiak, M. Szczerska, M. Wróbel, A. Łuczkiewicz, S. Fudala-Książek, P. Wityk. Optical method supported by machine learning for detection of urinary tract infection and assessment of risk for urosepsis. Journal of Biophotonics, 16(8):e202300095, 2023. https://onlinelibrary.wiley.com/doi/full/10.1002/jbio.202300095.

[2] F. Gurcan and A. Soylu. Learning from imbalanced data: Integration of advanced resampling techniques and machine learning models for enhanced cancer diagnosis and prognosis. Cancers, 16(19):3417, 2024. https://pmc.ncbi.nlm.nih.gov/articles/PMC11476323/.

[3] V. Thambawita, P. Salehi, S.A. Sheshkal, S.A. Hicks, H.L. Hammer, S. Parasa, T. de Lange, P. Halvorsen, and M.A. Riegler. SinGAN-Seg: Synthetic training data generation for medical image segmentation. PLOS ONE, 17(5):e0267976, 2022. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0267976.

[4] S. Szeghalmy and A. Fazekas. A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. Sensors, 23(4):2333, 2023. https://pmc.ncbi.nlm.nih.gov/articles/PMC9967638/.