SHARPNESS-AWARE MINIMIZATION (SAM) IMPROVES CLASSIFICATION ACCURACY OF BACTERIAL RAMAN SPECTRAL DATA ENABLING PORTABLE DIAGNOSTICS

Kaitlin Zareno*, **Jarett Dewbury***, **Loza F. Tadesse**
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{kzareno,jdewbury,lozat}@mit.edu

Siamak K. Sorooshyari
Department of Statistics
Stanford University
Palo Alto, CA 94305, USA
siamak@stanford.edu

Hossein Mobahi
Google Research
Mountain View, CA 94043, USA
hmobahi@google.com

ABSTRACT

Antimicrobial resistance is expected to claim 10 million lives per year by 2050, and resource-limited regions are most affected. Raman spectroscopy is a novel pathogen diagnostic approach promising rapid and portable antibiotic resistance testing within a few hours, compared to days when using gold standard methods. However, current algorithms for Raman spectra analysis 1) are unable to generalize well on limited datasets across diverse patient populations and 2) require increased complexity due to the necessity of non-trivial pre-processing steps, such as feature extraction, which are essential to mitigate the low-quality nature of Raman spectral data. In this work, we address these limitations using Sharpness-Aware Minimization (SAM) to enhance model generalization across a diverse array of hyperparameters in clinical bacterial isolate classification tasks. We demonstrate that SAM achieves accuracy improvements of up to 10.5% on a single split, and an increase in average accuracy of 2.7% across all splits in spectral classification tasks over the traditional optimizer, Adam. These results display the capability of SAM to advance the clinical application of AI-powered Raman spectroscopy tools. Code is available at: https://github.com/Tadesse-Lab/SAM-Raman-Diagnostics

1 INTRODUCTION

Antimicrobial resistance is the second biggest global health threat and is expected to surpass cancer by 2050 (de Kraker et al., 2016) as the second leading cause of death in the United States. According to the Centers for Disease Control and Prevention (CDC), nearly 50% of all outpatient antibiotic usage is improper, including unnecessary use and inappropriate selection, dosing, and duration of antibiotic treatments (Centers for Disease Control and Prevention (CDC), 2011; Pichichero, 2002; Shapiro et al., 2014). Altogether, such usage compromises effectiveness and contributes significantly to antimicrobial resistance. In the context of this escalating challenge, it is crucial to recognize that certain regions, such as developing nations or conflict zones, where access to quality hospital-level clinical care is limited, are most vulnerable. Thus, a rapid, reliable, and compact solution for bacterial infection diagnosis is needed to provide timely and accurate treatment. A promising tool is Raman spectroscopy, an optical technique that captures inelastically scattered light acting as a fingerprint for precise mapping to specific pathogen types. Within seconds, this method can identify not only the primary pathogen but also its mutated antibiotic resistant variants (Das & Agrawal, 2011; Garcia-Rico et al., 2018; Pieczonka & Aroca, 2008).

In this work, we address the intrinsic data quality challenges of working with Raman spectral data by utilizing Sharpness-Aware Minimization (SAM) as an optimizer for Raman spectral data analysis.

Our main contribution is as follows: we demonstrate that SAM is an effective optimizer for clinical Raman spectroscopy-based pathogen classification tasks, achieving an increase in accuracy of up to 10.5% on a single split, and an increase in average accuracy of 2.7% across all splits when compared to a standard ResNet architecture with Adam. The results demonstrate accuracy and generalizability in translating rapid on-device pathogen diagnostics for patients in need.

2 Background and Related Work

Although Raman spectroscopy is a promising tool for portable and rapid infectious disease diagnostics, Raman spectral data tends to be complex, high-dimensional, and noisy. Due to intrinsic limitations in the quality of Raman spectral data, current ML-based (machine learning) approaches in Raman spectroscopy often encounter 1) increased complexity due to the necessity of extra enhancing steps, such as feature extraction (Gautam et al., 2015; Pelletier, 2003), and 2) poor generalization due to limited datasets and small sample sizes, which have high variability both within individual patient samples and across patient populations (Luo et al., 2022). These data challenges limit the clinical translation of portable Raman spectroscopy-based diagnostics.

Deep learning offers the advantage of extracting non-linearities in the data without the need for non-trivial pre-processing (Liu et al., 2017), allowing for improved classification and more rapid processing in clinical translation. 1D convolutional neural networks (CNNs) and ResNets are the most commonly applied models in Raman classification tasks, and certain architectures (Figure 1) have achieved accuracies of up to 99% in classifying pathogen types and antibiotic resistance using Raman spectra (Ho et al., 2019; Ogunlade et al., 2023). However, prior ML classification tasks have required multiple samples to make a majority class prediction, and since sampling is a time-consuming process, it is less desirable for clinical translation where rapid assessment is needed.

SAM offers a potential solution to the generalization challenges that come with the application of traditional ML approaches to small datasets. Unlike previous optimization methods that may develop sharp minima and have poor generalization, SAM is an optimization technique that improves generalization by minimizing both the loss value and loss sharpness simultaneously (Foret et al., 2021). The ability to effectively generalize to unseen data makes SAM a useful tool for addressing the limited data issue commonly encountered in deep learning inference tasks with Raman spectral data. Despite its success in healthcare-related tasks (Song et al., 2022; Anand et al., 2022) and its promise towards advancing generalization in tasks with data quality concerns, SAM has not yet been applied to Raman spectral data. In this work, we address this gap by using SAM to enhance model generalization across a diverse array of hyperparameters in patient-derived bacterial classification tasks.

3 Methods

3.1 Sampling Method:

Clinical significance was preserved through subject-level sampling by excluding complete patient samples from either the train or test set. This approach ensured that our model had not seen any spectra from the clinical isolates in the test set during training, maintaining the integrity of the simulation on a subject-level. We utilized this method in our clinical evaluation pipeline as described below.

To demonstrate our algorithm's performance, we used publicly available spectral data from a leading Raman spectroscopy and CNN-based work by Ho et al. (2019). We obtained clinical bacterial isolate spectral data spanning the five most common first-line antibiotic treatment groups. This data was collected in groups of two distinct biological replicates. The dataset comprises 50 distinct clinical isolates collected from individual patients, categorized into five bacterial pathogen classes. It includes an equal distribution of five patient sources per infection type. To assess the efficacy of SAM with a ResNet architecture for Raman we performed random stratified splitting on this clinical spectral dataset.

To demonstrate clinically relevant evaluation, we began by randomly assigning one patient per infection type from each of the two biological replicates to the test set (Figure A.1). The remaining

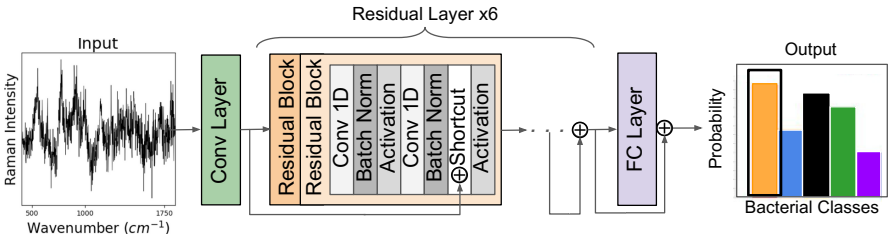


Figure 1: A CNN trained on clinical Raman spectral data is used for a pathogen classification task. Publicly available spectral data and CNN architecture are adapted from Ho et al. (2019). Using a one-dimensional residual network with 6 residual blocks and 25 total convolutional layers, clinical Raman spectra are classified as one of 5 isolates.

four patients in each type were utilized for training where they were split 90/10 into train and validation sets. We then evaluated each model on the test dataset which contains independent clinical data that has not been seen in the training pipeline. We repeated this process 10 times, employing randomized splits for each iteration (Figure A.1b). This approach was utilized to enhance variation within test sets and introduce a diverse set of data splits. We report classification results on the test dataset across five trials per split.

3.2 Architecture Configuration and Select Hyperparameters

For our baseline architecture, we utilized the ResNet described in Ho et al. (2019) (Figure 1) due to the architecture's proven efficacy in bacterial classification tasks. Throughout the course of experimentation, this architecture and the associated hyperparameters were modified in order to optimize model performance. During experimentation, we examined the use of Gaussian Error Linear Unit (GELU) activation (Hendrycks & Gimpel, 2016), Rectified Linear Unit (ReLU) activation (Nair & Hinton, 2010), and Scaled Exponential Linear Unit (SELU) activation (Klambauer et al., 2017). We trained all models on a single host having 1 NVIDIA T4 GPU.

3.3 SAM Improves Generalization by Finding Flat Minima

For our baseline optimizer, we used vanilla Adam (Kingma & Ba, 2015), a common choice in deep learning. While Adam and similar optimizers focus on finding parameters with low loss values, the loss surface geometry (particularly its sharpness around optimal solutions) has been shown to significantly affect model generalization (Keskar et al., 2016). This observation led to the development of Sharpness-Aware Minimization (SAM), which seeks parameters in regions of uniformly low loss Foret et al. (2021). We utilized SAM as described in Foret et al. (2021) to find parameters $\boldsymbol{w}$ within flat loss basins by optimizing the following objective:

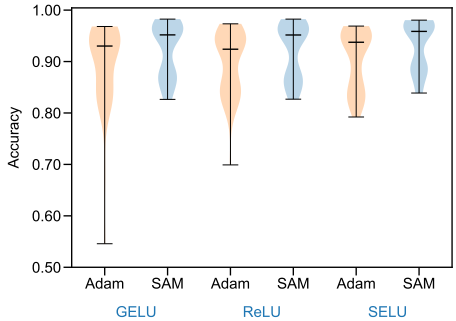
$$\begin{aligned} \min_{\boldsymbol{w}} \quad & L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_2^2 \\ \text{where} \quad & L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) = \max_{\|\boldsymbol{\epsilon}\|_p \leq \rho} L_{\mathcal{S}}(\boldsymbol{w} + \boldsymbol{\epsilon}) \end{aligned} \tag{1}$$

To minimize $L_{\mathcal{S}}^{SAM}(\boldsymbol{w})$, the inner maximization was linearized and solved analytically so that the gradient $\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w})$ could be effectively approximated as follows:

$$\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{SAM}(\boldsymbol{w}) \approx \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(w)|_{\boldsymbol{w}+\hat{\boldsymbol{\epsilon}}(\boldsymbol{w})} \tag{2}$$

The model is more likely to generalize effectively and resist perturbations as the resulting parameter values which exist in neighborhoods with uniformly low loss result in minimal variation in loss relative to the training objective function (Keskar et al., 2016). Thus, by finding regions of uniformly low loss, SAM's optimization algorithm was shown to improve model generalization performance.

(a)



(b)

Optim.	GELU	ReLU	SELU
Adam	$89.3_{8.0}$	$89.7_{6.1}$	$90.4_{5.7}$
SAM	$92.3_{5.2}$	$92.1_{5.4}$	$93.1_{4.8}$

(c)

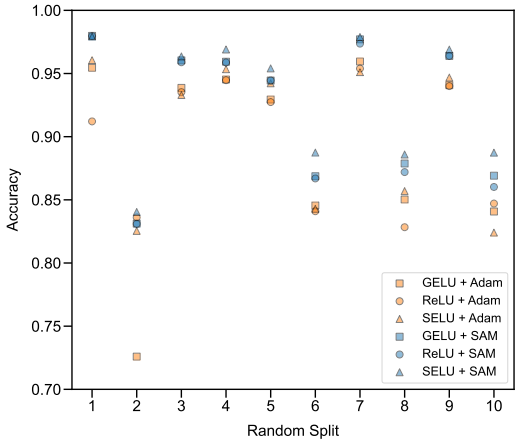


Figure 2: Performance of SAM relative to Adam across splits and hyperparameters. (a) A plot visualizing classification accuracy for each model, averaged across all 10 splits. We highlight SAM's performance, indicating higher accuracy and lower variance across hyperparameters. (b) A table reporting the average classification accuracy of each configuration, showing that SAM increases average accuracy relative to Adam across hyperparameters. SAM with SELU has the best overall performance with an average accuracy of $93.1 \pm 4.8\%$. (c) A summary of results for each individual split demonstrates that ResNet configurations with SAM (blue) tend to outperform those using Adam (orange) across all splits.

This makes it ideal for limited, high-variance Raman spectra derived from clinical samples. For our experiments, we utilized a PyTorch implementation of SAM [1].

4 Results and Discussion

Given the implemented ResNet architecture and hyperparameters, we observed that the SAM optimizer increased average accuracy up to 3.0% when compared to Adam, as shown in Figure 2b. Average performance was determined for each configuration by averaging across all of the 10 splits and 5 trials. Thus, the results reflect the expected model performance and inherent variances within the test dataset. Table 1 presents a detailed summary of the results across all evaluated splits. As illustrated in Figure 2c, individual data splits highlight SAM's performance; when compared to Adam, SAM demonstrated a maximum increase in performance of 10.5% on a single split. SAM consistently outperformed Adam in mean classification accuracy across all splits. For each split, GELU experienced, on average, an increase in accuracy of 3.0% with SAM compared to Adam. Similarly, ReLU and SELU achieved 2.4% and 2.8% increases in average accuracy, respectively. In

[1] https://github.com/davda54/sam

Table 1: Testing set performance ($mean_{std}$ for n=5) across different splits and hyperparameters trained with Adam and SAM optimizers. Training with SAM yields higher classification accuracy across most of the splits and hyperparameters when compared to Adam.

	Optim.	Random Splits 1	2	3	4	5	6	7	8	9	10
GELU	Adam	$95.5_{1.2}$	$72.6_{10.2}$	$93.9_{0.7}$	$94.5_{0.9}$	$92.9_{0.4}$	$84.6_{1.2}$	$96.0_{0.7}$	$85.0_{2.5}$	$94.1_{1.1}$	$84.1_{1.3}$
	SAM	$98.0_{0.2}$	$83.1_{0.4}$	$96.0_{0.3}$	$95.9_{0.1}$	$94.4_{0.2}$	$86.9_{0.3}$	$97.7_{0.1}$	$87.9_{0.5}$	$96.4_{0.2}$	$86.9_{0.3}$
ReLU	Adam	$91.2_{10.7}$	$83.6_{2.2}$	$93.5_{0.8}$	$94.4_{0.6}$	$92.7_{1.0}$	$84.1_{1.5}$	$95.4_{1.6}$	$82.8_{2.4}$	$94.0_{2.1}$	$84.7_{0.8}$
	SAM	$98.0_{0.2}$	$83.1_{0.4}$	$95.9_{0.1}$	$95.9_{0.2}$	$94.5_{0.4}$	$86.7_{0.6}$	$97.4_{0.1}$	$87.2_{0.2}$	$96.4_{0.1}$	$86.0_{0.5}$
SELU	Adam	$96.1_{1.0}$	$82.5_{1.0}$	$93.3_{2.2}$	$95.3_{1.7}$	$94.2_{0.4}$	$84.3_{0.8}$	$95.1_{0.7}$	$85.7_{2.8}$	$94.7_{1.4}$	$82.4_{1.9}$
	SAM	$97.9_{0.1}$	$84.0_{0.1}$	$96.4_{0.2}$	$96.9_{0.2}$	$95.4_{0.1}$	$88.7_{0.3}$	$97.9_{0.2}$	$88.6_{0.3}$	$96.9_{0.1}$	$88.7_{0.2}$

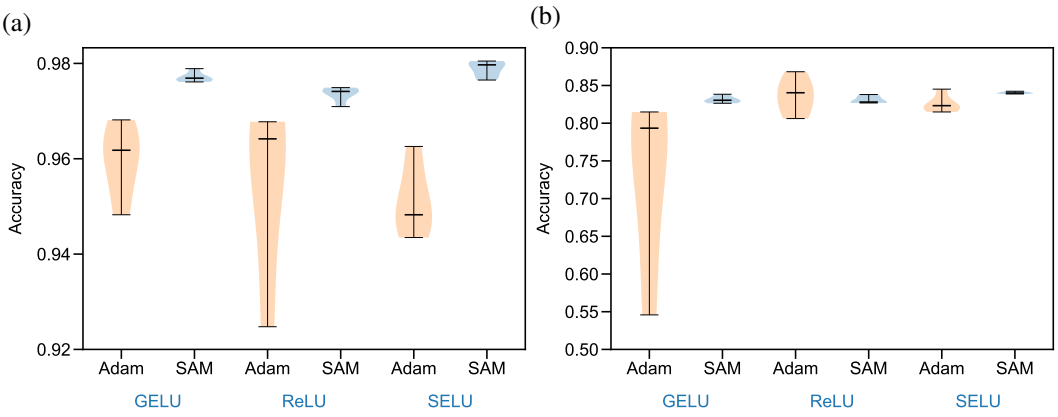


Figure 3: Classification accuracy of each technique on the best and worst performing splits across all hyperparameters. (a) ResNet trained with SAM significantly outperforms Adam across all hyperparameters in the best performing split. (b) Training with SAM has lower variance but no significant performance improvement over Adam in the worst performing split.

addition, across all splits, SAM considerably decreased the expected variance in accuracy. When using GELU, ReLU, and SELU, SAM also demonstrated decreases in variance of 79.1%, 75.2%, and 84.3% on average, respectively, as shown in Table 1.

In addition to looking at the splits as a whole, we observed the best and worst performing splits to better understand how SAM performed at the limits of our test sets (Figure 3). The best and worst performing splits were determined by observing the mean performance across all splits. In the best performing split (Figure 3a), we found that SAM significantly outperformed Adam across all hyperparameters, offering both increased accuracy and decreased performance variance. In the worst performing split (Figure 3b), we observed that SAM provided no significant performance improvement compared to Adam. Figure A.2 further details the performance of SAM and Adam across the remaining splits. Although overlapping regions of standard deviations (Figure 2a) are observed in average performance across all splits, the lower variance and statistically significant improved accuracy performance in the best performing cases suggest that SAM is positioned to be a more promising approach for clinical Raman spectral classification tasks. The increase in accuracy observed when using SAM as opposed to Adam indicates the promise of SAM for spectral data classification tasks. Furthermore, in clinical classification tasks, consistency in prediction is as crucial as performance itself. SAM exhibited a 2.7% increase in average performance and a 79.5% reduction in average split-wise variance across activations, suggesting a more stable model and providing greater reliability in prediction for classification tasks where the ability to validate results is limited.

5 Conclusion

In this work, we developed solutions for addressing limitations associated with limited datasets in novel Raman spectroscopy-based rapid infection diagnostics and antibiotic susceptibility testing tools. Compared to the standard utilization of the Adam optimizer for Raman spectral analysis, training with SAM demonstrated an increase in accuracy of up to 10.5% on a single split, an average accuracy improvement of 2.7% across all splits, and considerable reduction in variance across the chosen hyperparameters. Performance with SAM was consistent across 10 distinct splits, showcasing its robustness in processing a variety of patient-derived spectral data. Furthermore, the results indicate a potential to better classify diverse bacterial profiles, suggesting that SAM may provide a more reliable approach for clinical single-spectrum bacterial classification. Our work contributes to the translation of portable Raman spectroscopy-based diagnostics towards the fight against the global health threat of antibiotic resistance.

Limitations. We recognize that further exploration is required to determine the viability of SAM across different datasets and architectures. Furthermore, we acknowledge that the clinical dataset employed in this study may not comprehensively represent patient demographics and may not account for the variance encountered when collecting spectra samples from different Raman instruments.

ACKNOWLEDGMENTS

We acknowledge support from the MIT-Google Program for Computing Innovation.

REFERENCES

Deepa Anand, Rohan Patil, Utkarsh Agrawal, Rahul V, Hariharan Ravishankar, and Prasad Sudhakar. Towards generalization of medical imaging ai models: Sharpness-aware minimizers and beyond. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2022. doi: 10.1109/ISBI52829.2022.9761677.

Centers for Disease Control and Prevention (CDC). Office-related antibiotic prescribing for persons aged $\leq$ 14 years–united states, 1993-1994 to 2007-2008. *MMWR Morb. Mortal. Wkly. Rep.*, 60 (34):1153–1156, September 2011.

Ruchita S. Das and Y.K. Agrawal. Raman spectroscopy: Recent advancements, techniques and applications. *Vibrational Spectroscopy*, 57(2):163–176, 2011. ISSN 0924-2031.

M. E. de Kraker, A. J. Stewardson, and S. Harbarth. Will 10 million people die a year due to antimicrobial resistance by 2050? *PLOS Medicine*, 13(11):e1002184, November 2016. ISSN 1549-1676.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Eduardo Garcia-Rico, Ramon A Alvarez-Puebla, and Luca Guerrini. Direct surface-enhanced raman scattering (SERS) spectroscopy of nucleic acids: from fundamental studies to real-life applications. *Chem. Soc. Rev.*, 47(13):4909–4923, 2018.

Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for raman and infrared spectroscopy. *EPJ Tech. Instrum.*, 2(1), December 2015.

Dan Hendrycks and Kevin Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. *CoRR*, abs/1606.08415, 2016. URL http://arxiv.org/abs/1606.08415.

Chi-Sing Ho, Neal Jean, Catherine A Hogan, Lena Blackmon, Stefanie S Jeffrey, Mark Holodniy, Niaz Banaei, Amr A E Saleh, Stefano Ermon, and Jennifer Dionne. Rapid identification of pathogenic bacteria using raman spectroscopy and deep learning. *Nat. Commun.*, 10(1):4927, October 2019.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *CoRR*, abs/1609.04836, 2016. URL http://arxiv.org/abs/1609.04836.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *ICLR (Poster)*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017. URL http://arxiv.org/abs/1706.02515.

Jinchao Liu, Margarita Osadchy, Lorna Ashton, Michael Foster, Christopher J Solomon, and Stuart J Gibson. Deep convolutional neural networks for raman spectrum recognition: a unified solution. *Analyst*, 142(21):4067–4074, 2017.

Ruihao Luo, Juergen Popp, and Thomas Bocklitz. Deep learning for raman spectroscopy: A review. *Analytica*, 3(3):287–301, 2022. ISSN 2673-4532. doi: 10.3390/analytica3030020. URL https://www.mdpi.com/2673-4532/3/3/20.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *International Conference on Machine Learning*, 2010. URL https://api.semanticscholar.org/CorpusID:15539264.

Babatunde Ogunlade, Loza F Tadesse, Hongquan Li, Nhat Vu, Niaz Banaei, Amy K Barczak, Amr A E Saleh, Manu Prakash, and Jennifer A Dionne. Predicting tuberculosis drug resistance with machine learning-assisted raman spectroscopy. *arXiv*, arXiv:2306.05653v1, June 2023. URL http://arxiv.org/abs/2306.05653v1.

M J Pelletier. Quantitative analysis using raman spectrometry. *Appl. Spectrosc.*, 57(1):20A–42A, January 2003. doi: 10.1366/000370203321165133.

Michael E. Pichichero. Dynamics of Antibiotic Prescribing for Children. *JAMA*, 287(23):3133–3135, 06 2002. ISSN 0098-7484. doi: 10.1001/jama.287.23.3133. URL https://doi.org/10.1001/jama.287.23.3133.

Nicholas P. W. Pieczonka and Ricardo F. Aroca. Single molecule analysis by surfaced-enhanced raman scattering. *Chem. Soc. Rev.*, 37:946–954, 2008. doi: 10.1039/B709739P. URL http://dx.doi.org/10.1039/B709739P.

Daniel J Shapiro, Lauri A Hicks, Andrew T Pavia, and Adam L Hersh. Antibiotic prescribing for adults in ambulatory care in the USA, 2007-09. *J. Antimicrob. Chemother.*, 69(1):234–240, January 2014.

Yeong-Hun Song, Jun-Young Yi, Young Noh, Hyemin Jang, Sang Won Seo, Duk L Na, and Joon-Kyung Seong. On the reliability of deep learning-based classification for alzheimer's disease: Multi-cohorts, multi-vendors, multi-protocols, and head-to-head validation. *Front. Neurosci.*, 16: 851871, September 2022.

A APPENDIX

A.1 EXPERIMENTAL METHODOLOGIES AND RESULTS

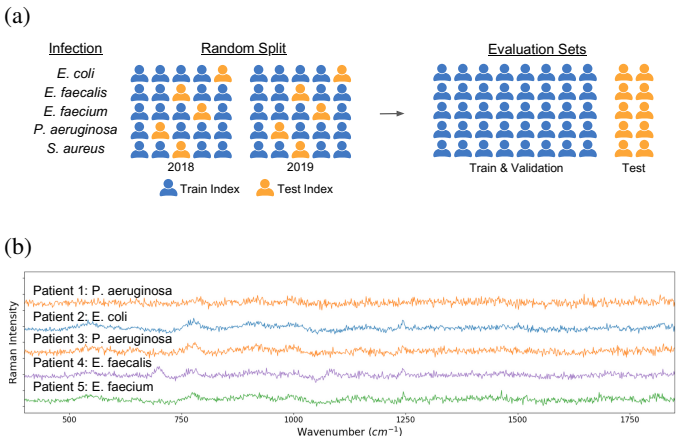


Figure A.1: A CNN trained on clinical Raman spectral data is utilized for a pathogen classification task. Publicly available spectral data and CNN architecture are adapted from Ho et al. (2019). (a) Two datasets of 5 patients per infection type and 5 species of bacterial infections are merged to create a single dataset of 5 species of bacterial infection and 10 patients per infection type. Each patient is classified into one of the 5 treatment classes where each species corresponds to a different treatment class. Average species identification accuracy using SAM shows an improvement from $89.8 \pm 6.7\%$ to $92.5 \pm 5.2\%$. (b) Noisy spectra and the intrinsic qualities of Raman spectral data can make it difficult to distinguish between species, highlighting the need for deep learning and SAM to enable rapid analysis of Raman spectral data.

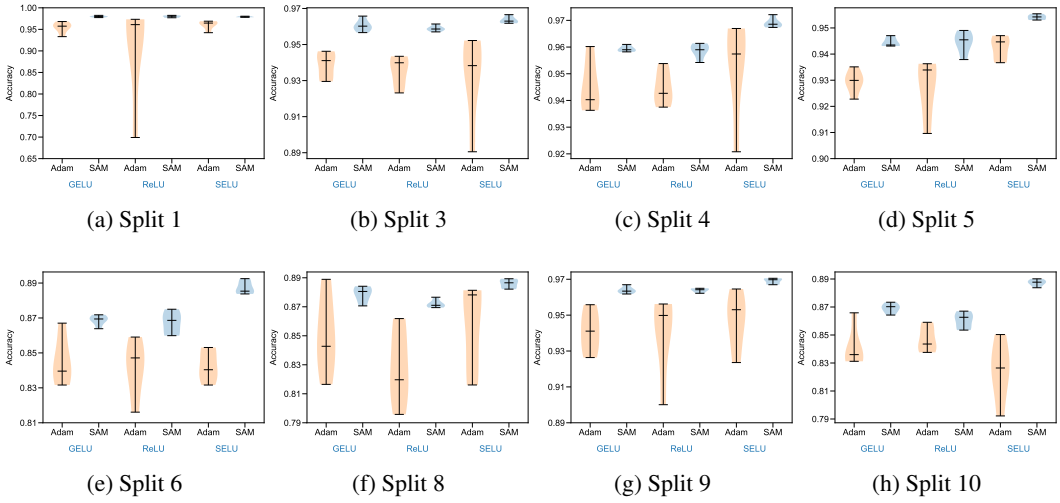


Figure A.2: Individual plots showing classification performance $mean_{std}$ (n=5) for evaluation metrics on the testing set across the reported random splits.