# S-MSRRS5000: A Simulated Dataset Highlighting the Challenges of Data Obtained From Multiple Spatially Resolved Reflection Spectroscopy

Birk Martin Magnussen*, Maik Jessulat†, Claudius Stern‡ and Bernhard Sick†
*biozoom services GmbH
34121 Kassel, Germany
†Intelligent Embedded Systems
Universität Kassel
34121 Kassel, Germany
‡FOM Hochschule für Oekonomie & Management
Hochschulzentrum Kassel
34117 Kassel, Germany
e-mail: birk.magnussen@biozoom.net, {mjessulat, bsick}@uni-kassel.de, Claudius.Stern@fom.de

***Abstract*—Optical sensors based on spectroscopy are occasionally used in consumer healthcare and wellness applications. This includes applications such as measuring the concentration of cutaneous carotenoids using sensors based on multiple spatially resolved reflection spectroscopy (MSRRS). Processing the data yielded from MSRRS-based sensors poses unique challenges. When using machine learning for data processing, specialized models such as continuous feature networks are required to achieve good results. However, due to privacy issues of medical data, data availability is low, hindering model development. In this article, a simulated dataset is introduced, highlighting the challenges of data from MSRRS-based sensors. Furthermore, the underlying principles used for the simulation will be discussed. Finally, several model architectures including continuous feature networks are trained on the dataset, demonstrating the various challenges.**

***Index Terms*—dataset, simulated data, reflection spectroscopy**

## I. Introduction

Multiple spatially resolved reflection spectroscopy (MSRRS) is a technology for analyzing biological tissue. One common use-case for sensors based on MSRRS is measuring the concentration of cutaneous carotenoids in humans [1]. MSRRS-based sensors consist of multiple light emitters and light detectors. These emitters and detectors are then arranged in a matrix. When a measuring sample is placed on this detector matrix, the light emitted from an emitter will pass through the sample before being detected by the light detector. The measured brightness at the light detector then depends on the wavelength of the light, the distance between emitter and detector, as well as the optical properties of the sample. By analyzing the brightness observed for all emitter-detector pairs, it is possible to make predictions about the measurement sample [2], such as the concentration of cutaneous carotenoids.

This analysis of the observed brightness values can be performed using machine learning. For this purpose, specialized neural network architectures have been developed, such as continuous feature networks [3], [4]. These continuous feature networks are designed to tackle the difficulties of data yielded by MSRRS-based sensors.

At the core of the difficulty of processing MSRRS-based data is the irregularity of the data. The light emitters used in MSRRS-based sensors have discrete wavelengths. As a result, the observed spectrum is not sampled continuously, but rather at discrete sampling points. Furthermore, these points are not distributed equally but are at irregular intervals. A

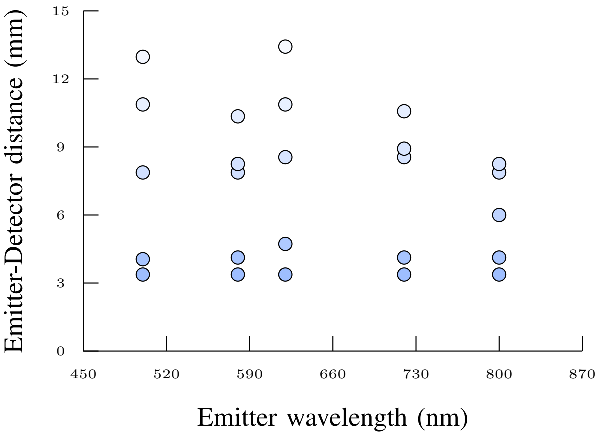

Fig. 1. An example of the structure of data from MSRRS-based sensors [4].

Supported by the state Hessen through funding as part of the Distr@l research grant 22_0041_2B.

similar situation is observed with the distance of emitter-detector pairs. An example of this type of irregular data can be seen in Figure 1.

In addition to the irregularity, any MSRRS-based sensor can be afflicted by manufacturing inaccuracies. This can result in the wavelengths of the light emitters shifting slightly between individual sensors. Any machine learning model operating on data from MSRRS-based sensors needs to be able to compensate for these fluctuations.

When trying to predict properties or substances in human tissue, the required training data is composed of medical data. Due to privacy concerns, such data cannot be made publicly available. Therefore, a simulated dataset called S-MSRRS5000 is presented.

## II. The Dataset and its Challenges

For the creation of S-MSRRS5000, a virtual MSRRS-based sensor with 8 detectors and 32 emitters is assumed.

S-MSRRS5000 is made up of four sets of 5000 simulated measurements each. Each measurement consists of a series of 256 data points. One data point always corresponds to one emitter-detector pair of the virtual sensor. For each data point, the observed brightness at the detector is listed. In addition, the wavelength of the emitter and the emitter-detector distance are listed. Together, the 256 data points make up one measurement. For each such measurement, three ground truth labels are available.

1) The concentration of cutaneous beta-carotene in the simulated sample in milligrams per liter.
2) The concentration of hemoglobin in the blood of the sample in grams per liter.
3) The oxygenation level of the blood in the sample as a factor from 0 to 1.

The ground truth labels for each set are stored in a file called `dataset_meta.json`. This JSON file contains a single top-level array. Within are multiple JSON objects, each representing one simulated measurement, with a key for each label and a key called `name` containing the filename of the measurement data itself. The file containing the measurement data is a CSV file with a list of data points. Each data point is listed with the wavelength of the emitted light, the emitter-detector distance, and the measured intensity.

The four sets available in the dataset represent different challenges of MSRRS-based data in ascending difficulty. The challenges of the different sets are as follows:

1) **`1_clean`: Clean Data Only** The first set represents the ideal case. In this set, the virtual MSRRS-based sensor is perfectly accurate. Similarly, the measurement sample is simulated to have no noise or other factors negatively impacting the data. The only factors with a variable influence on the simulated brightness values are substances with labels available as ground truth. This set is thus useful to understand the structure of MSRRS-based data and to establish a baseline for model performance.
2) **`2_noise`: Noisy Samples** The second set increases the prediction difficulty by introducing measurement noise. In addition to the substances with labels available as ground truth, further randomized noise contributes to the light absorption within the simulated sample. Furthermore, the carotenoid concentration is simulated with localized fluctuations and is no longer homogeneous throughout the entirety of the sample.
3) **`3_wavelength_shift`: Inconsistent Emitter Wavelengths** The third set introduces production inaccuracies of the sensor in addition to the noise of the second set. In this set, the emitter wavelengths of each measurement are no longer identical. Instead, they are simulated to fluctuate in a small region around the nominal target wavelength that is used in the previous sets.
4) **`4_missing_data`: Missing Data From Detectors** The fourth set expands on the difficulty of the third set. In addition to the sample noise and inconsistent emitter wavelength, the fourth set allows for inoperable detectors. As a result, it is no longer guaranteed that all data points are available in a measurement. Instead, some data points may be replaced by NaN-values. This is to simulate detectors that might have to be disabled due to them exceeding tolerances or being otherwise defective.

## III. Simulation Background

The simulation of the brightness detected by the light detectors is a highly simplified simulation. It does not aim to accurately recreate the measurement values as when measuring real human tissue but rather aims at recreating some of the core challenges of real MSRRS-based data.

At its core, the simulation traces the light paths from every light emitter to every light detector as semicircles. These semicircles are split into equidistant segments. Figure 2 shows an example of these light paths for two emitters and one detector.

Starting from the light emitter, the light attenuation over the light path segments is calculated. In this simulation, attenuation is comprised of two components. First, since the light emitter is not emitting a focused beam of light, the light attenuates based on the distance to the emitter with the inverse-square law. Given the distance to the light emitter $d$ and the length of the light path segment $l_s$, the ratio between the light intensity $I_0$ before the light path segment and the light

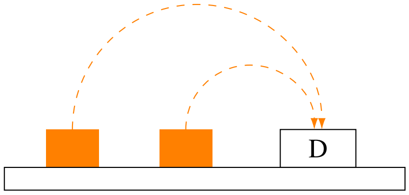

Fig. 2. Traced light paths split into equidistant segments of two emitters to one light detector.

intensity $I$ after the light path segment can be described as follows:

$$\frac{I}{I_0} = \frac{d^2}{\left(d + l_s\right)^2} \tag{1}$$

The second component is the absorption of light by the sample, calculated using the Beer-Lambert law. The Beer-Lambert law describes the light attenuation due to the absorbance of light of the chemicals. Given the absorptivity $a$ of a chemical, optical beam length $b$, and the concentration of the absorbing chemical $c$, the Beer-Lambert law allows to calculate the logarithmic ratio between the light intensity $I_0$ before the light path segment and the light intensity $I$ after the light path segment can be described as follows [5]:

$$-\log\left(\frac{I}{I_0}\right) = a \cdot b \cdot c \tag{2}$$

The absorptivity $a$ is a wavelength-dependent material constant. The optical path length $b$ can be calculated as the product of the light path segment length and the refractive index of human skin (approximated as $1.37$ [6]).

This dataset simulates the to-be-measured tissue as three primary components. First, the tissue is assumed to consist of a base of water. Additionally, the tissue contains hemoglobin as found in blood, both oxygenated and not oxygenated. Finally, a given concentration of carotenoids contributes to the absorption of light.

For the base of water, the absorptivity used by the simulation is based on the measurements of Kou, Labrie, and Chylek [7].

For the hemoglobin contained in the blood within the simulated tissue, measurements compiled by Prahl [8] were used as the absorptivity values in the simulation. The amount of hemoglobin present for absorption is calculated as a combination of the amount of hemoglobin per volume of blood, and the ratio of volume of blood to volume of total tissue. The volume of blood present in human limbs averages $5.6\%$ of the tissue volume [9]. For the concentration of hemoglobin in human blood, values are sampled from a distribution based on measurements from Kim et al. [10] and stored as labels in the ground truth data. Similarly, for the ratio of oxygenated to non-oxygenated blood, a value is sampled from a distribution based on measurements from Epstein and Haghenbeck [11] and similarly stored as a label in the ground truth data.

For the carotenoids to be predicted by a given machine learning algorithm, a concentration is sampled from a distribution of the beta-carotene concentrations in humans based on measurements from Matsumoto et al. [12] and stored as a label in the ground truth data. Data compiled by Prahl [13] on the absorptivity spectrum of beta-carotene was used as the basis for the simulation.

For the noise in set two and beyond, OpenSimplex-based noise [14] is used. A three-dimensional noise map is used to vary the localized concentration of carotenoids by up to $\pm 15\%$. A second, four-dimensional is used as well. The first three dimensions represent the position in the simulated space while the fourth dimension represents the current wavelength of light. This noise map is used to add localized and wavelength-dependent background absorptivity noise, representing other substances present in the tissue.

The light emitter wavelength shift in the third and fourth sets is sampled from a Gaussian distribution around the target wavelength for the light emitter with a standard deviation of $2.5$ nm.

In the fourth set, the chance that any single detector is disabled for the simulation is fixed at $10\%$. This ratio is higher than expected for real MSRRS-based sensors but serves as a suitable worst-case scenario challenge for a prediction model.

## IV. Example Usage and Results

For testing, three different machine learning models were trained on each of the four sets within the S-MSRRS5000 dataset. The three models include a simple linear regression model, a multi-layer feed-forward neural network, and a continuous feature network as proposed in Magnussen, Stern, and Sick [3], [4]. For this experiment, each set of the dataset was split into 3500 training measurements, and 1500 validation measurements. For both methods based on neural networks, the 1500 validation measurements were further split into 500 testing measurements and only 1000 final validation measurements. The testing measurements were used to select the best-performing model during training, while the validation measurements are the basis for the final score of the model. The training was repeated ten times for each model and each set, with randomized training, testing, and validation data splits for each repetition.

The linear regression model and the multi-layer feed-forward network are not able to deal with missing input data for the fourth set in S-MSRRS5000. Instead, the missing values were imputed by linearly interpolating values of comparable wavelength and emitter-detector distances. The continuous feature network is capable of handling missing input data without imputation and was thus given the input data of the fourth set unmodified.

The results shown in Table I include the root mean square error (RMSE) and the coefficient of determination ($r^2$) between the predicted results and the ground truth labels of the dataset. The data is averaged over the ten runs, with the respective standard deviation given for each quality metric. It is to be noted that only training instances that yielded a positive coefficient of determination contributed to the respective quality metrics. Instead, the % fail metric indicates the percentage of runs that yielded a negative coefficient of determination, indicating models with an unusable training result.

From the data in Table I, it can be observed that calculating the concentration of carotenoids in a simplified simulation such as used for S-MSRRS5000 is a simple task if no noise is present. All models trained are able to achieve a very high coefficient of determination. However, especially the introduction of fluctuations in the wavelength of the light emitters has a significant impact on the linear regression model and

TABLE I

The training accuracy of different models on the various challenges of the S-MSRRS5000 dataset.

| | linear regression | | | multi-layer feed-forward network | | | continuous feature network | | |
|---|---|---|---|---|---|---|---|---|---|
| | RMSE | $r^2$ | % fail | RMSE | $r^2$ | % fail | RMSE | $r^2$ | % fail |
| clean data | **0.0**±0.0 | **1.0**±0.0 | 0% | 0.02±0.0 | **1.0**±0.0 | 0% | 0.03±0.0 | 0.99±0.0 | 0% |
| noisy data | **0.03**±0.0 | **0.99**±0.0 | 0% | 0.04±0.0 | 0.98±0.0 | 0% | 0.05±0.01 | 0.97±0.01 | 0% |
| wavelength shift | 0.26±0.0 | 0.29±0.01 | 40% | - | - | 100% | **0.11**±0.01 | **0.85**±0.03 | 0% |
| missing data | 0.27±0.01 | 0.15±0.01 | 60% | - | - | 100% | **0.16**±0.02 | **0.60**±0.08 | 0% |

the multi-layer feed-forward network. The multi-layer feed-forward network was not able to achieve a positive coefficient of determination for a single training repetition for the third and fourth sets. The continuous feature network however is capable of taking the exact wavelength and emitter-detector distance for each input brightness into account. As a result, the continuous feature network is able to keep a high coefficient of determination and a low root mean square error for the third set as well. A similar effect can be observed for the fourth set, where the missing data significantly reduces the accuracy of the linear regression model, whereas the continuous feature network is able to deal with the missing data and still keeps a comparatively high coefficient of determination.

These results, especially including the ability of the continuous feature network to achieve a high performance due to being able to compensate for inconsistent wavelengths and for missing data is consistent with recent findings on real data [3], [4].

## V. Conclusion

The S-MSRRS5000 dataset is a simulated dataset, optimized to show the structure and challenges of MSRRS-based data. This article discusses the contents of the S-MSRRS5000 dataset and the four challenge sets making up the dataset. Furthermore, the relation of each challenge set to the challenges of real data is highlighted.

This article also goes into detail on how the dataset is simulated. This includes the underlying physical principles, as well as the sources for the spectral data used.

Finally, the dataset is used to train different example models on each of the challenge sets. The results from these experiments are consistent with results on real data.

## References

[1] M. E. Darvin, B. Magnussen, J. Lademann, and W. Köcher, “Multiple spatially resolved reflection spectroscopy for in vivo determination of carotenoids in human skin and blood,” *Laser Physics Letters*, vol. 13, no. 9, p. 095601, Aug. 2016.

[2] B. Magnussen, C. Stern, and W. Köcher, “Vorrichtung und verfahren zur bestimmung einer konzentration in einer probe,” European Patent EP 3 013 217 B1, 2016.

[3] B. M. Magnussen, C. Stern, and B. Sick, “Utilizing continuous kernels for processing irregularly and inconsistently sampled data with position-dependent features,” in *Proceedings of The Nineteenth International Conference on Autonomic and Autonomous Systems*, C. Behn, Ed., IARIA. ThinkMind, Mar. 2023, pp. 49–53.

[4] B. M. Magnussen, C. Stern, and B. Sick, “Continuous feature networks: A novel method to process irregularly and inconsistently sampled data with position-dependent features,” *International Journal On Advances in Intelligent Systems*, vol. 16, no. 3&4, pp. 43–50, 2023.

[5] D. F. Swinehart, “The beer-lambert law,” *Journal of Chemical Education*, vol. 39, no. 7, p. 333, Jul. 1962.

[6] H. Ding, J. Q. Lu, W. A. Wooden, P. J. Kragel, and X.-H. Hu, “Refractive indices of human skin tissues at eight wavelengths and estimated dispersion relations between 300 and 1600 nm,” *Physics in Medicine & Biology*, vol. 51, no. 6, p. 1479, Mar. 2006.

[7] L. Kou, D. Labrie, and P. Chylek, “Refractive indices of water and ice in the 0.65- to 2.5-$\mu$m spectral range,” *Applied Optics*, vol. 32, no. 19, pp. 3531–3540, Jul. 1993.

[8] S. Prahl. (1999) Optical absorption of hemoglobin. [Online]. Available: https://omlc.org/spectra/hemoglobin/index.html

[9] L. M. Karpeles and R. L. Huff, “Blood volume of representative portions of the musculoskeletal system in man,” *Circulation Research*, vol. 3, no. 5, pp. 483–489, 1955.

[10] S. K. Kim, H. S. Kang, C. S. Kim, and Y. T. Kim, “The prevalence of anemia and iron depletion in the population aged 10 years or older,” *The Korean Journal of Hematology*, vol. 46, no. 3, pp. 196–199, 2011.

[11] C. D. Epstein and K. T. Haghenbeck, “Bedside assessment of tissue oxygen saturation monitoring in critically ill adults: An integrative review of the literature,” *Critical Care Research and Practice*, vol. 2014, May 2014.

[12] M. Matsumoto, H. Suganuma, S. Shimizu, H. Hayashi, K. Sawada, I. Tokuda, K. Ihara, and S. Nakaji, “Skin carotenoid level as an alternative marker of serum total carotenoid concentration and vegetable intake correlates with biomarkers of circulatory diseases and metabolic syndrome,” *Nutrients*, vol. 12, no. 6, 2020.

[13] S. Prahl. (2017) Beta-carotene. [Online]. Available: https://omlc.org/spectra/PhotochemCAD/html/041.html

[14] KdotJPG. (2019) Opensimplex 2. [Online]. Available: https://github.com/KdotJPG/OpenSimplex2