A Deep Learning Framework for Spectrophotometric Quantification of Key Microalgal Pigments Supplementary materials

Omar Bayomie[1,2,3+] **and Victor Pozzobon**[4] ✉

[1]The Advanced Centre for Biochemical Engineering, Department of Biochemical Engineering, University College London, London, UK
[2]SSPC, The Science Foundation Ireland Research Centre for Pharmaceuticals, University College Dublin, Belfield, Dublin, Ireland
[3]Institut des Systemes Intelligents et de Robotique, Sorbonne University, Paris 75005, France
[4]Université Paris-Saclay, CentraleSupélec, Laboratoire de Génie des Procédés et Matériaux, Centre Européen de Biotechnologie et de Bioéconomie (CEBB), 3 rue des Rouges Terres 51110 Pomacle, France
[+]Contact email: omar.bayomie.23@ucl.ac.uk
[1]Corresponding author: victor.pozzobon@centralesupelec.fr

Algorithm flow

```
Input:
Spectral dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{1131}$ and $y_i \in \mathbb{R}$.
Pigment type $p \in \{\text{Chlorophyll a}, \text{Chlorophyll b}, \text{Lutein}, \text{Violaxanthin}, \text{Zeaxanthin}\}$.
Optimization Trials $T_{max}$.

Algorithm selection:
Split $\mathcal{D}$ into $\mathcal{D}_{train}$ (80%) and $\mathcal{D}_{holdout}$ (20%).
if $p =$ Zeaxanthin then
    Architecture $A \leftarrow$ Zeaxanthin-specific architecture for low abundance
    Target transformation: $y' \leftarrow \ln(1+y)$
    Loss function $\mathcal{L} \leftarrow$ Log-Cosh
else
    Architecture $A \leftarrow$ Standard CNN
    Target transformation: None
    Loss function $\mathcal{L} \leftarrow$ MAPE
end if

Optuna Loop
for trial $t = 1$ to $T_{\max}$ do
    Sample hyperparameters $\theta_t \sim \Omega$ using TPE sampler.
    $CV_{score} \leftarrow 0$.
    Inner Loop: CV
    for fold $k = 1$ to $K$ do
        Split $\mathcal{D}_{train}$ into $\mathcal{D}_{train}^{k}$ and $\mathcal{D}_{val}^{k}$.
        Construct 1D-CNN model $f(x;\theta_t)$ using Adam optimizer.
        $\theta_t^* \leftarrow \arg\min_\theta \sum \mathcal{L}\big(y, f(x;\theta)\big)$        Early Stopping applied
        $Score_k \leftarrow \text{MAPE}\big(y_{val}, f(x_{val};\theta_t^*)\big)$
        if $Score_k$ is significantly worse than the median of prior trials then
            stop trial $t$ (pruned).
        end if
        $CV_{score} \leftarrow CV_{score} + Score_k$
    end for
Return the mean cross-validation value $\frac{CV_{score}}{K}$ to Optuna.
```

Baseline model detailed statistics

Table 1 the performance of each of the baseline machine learning models in terms of MAPE for 5 different shuffles of the dataset. Table 2 presents the detailed statistical comparison of the baseline machine learning models. Best model columns is to be understood as the best model of the pairwise comparison.

Pigment	Preprocessing	PLS	Ridge	SVR
Chlorophyll a	Centered Reduced	38.10 ± 4.16	33.83 ± 4.02	35.59 ± 1.63
Chlorophyll a	EMSC	37.09 ± 5.28	31.67 ± 2.54	36.09 ± 1.74
Chlorophyll a	MSC	38.76 ± 4.70	33.39 ± 1.53	34.85 ± 1.99
Chlorophyll a	Derivatives only	13.02 ± 2.14	11.91 ± 0.00	15.78 ± 0.00
Chlorophyll b	Centered Reduced	31.98 ± 0.72	28.85 ± 1.10	25.39 ± 0.60
Chlorophyll b	EMSC	28.80 ± 3.03	26.62 ± 0.00	24.12 ± 0.00
Chlorophyll b	MSC	28.68 ± 0.99	28.66 ± 2.41	25.15 ± 0.70
Chlorophyll b	Derivatives only	10.04 ± 4.31	8.10 ± 1.96	8.38 ± 0.00
Lutein	Centered Reduced	22.85 ± 5.01	22.95 ± 2.97	19.99 ± 0.13
Lutein	EMSC	23.53 ± 3.21	21.58 ± 0.00	21.03 ± 0.00
Lutein	MSC	23.08 ± 5.38	22.51 ± 3.75	20.07 ± 0.26
Lutein	Derivatives only	9.21 ± 0.15	9.76 ± 0.00	6.54 ± 1.02
Violaxanthin	Centered Reduced	46.31 ± 3.11	41.12 ± 8.41	25.00 ± 0.47
Violaxanthin	EMSC	62.81 ± 9.94	55.52 ± 0.46	25.23 ± 1.88
Violaxanthin	MSC	44.70 ± 3.21	41.52 ± 0.00	24.42 ± 0.78
Violaxanthin	Derivatives only	15.52 ± 0.90	14.13 ± 1.29	16.59 ± 0.50
Zeaxanthin	Centered Reduced	32.54 ± 2.26	31.91 ± 0.00	17.58 ± 0.00
Zeaxanthin	EMSC	35.17 ± 2.00	33.60 ± 1.44	16.56 ± 0.14
Zeaxanthin	MSC	33.58 ± 2.42	32.42 ± 1.50	17.63 ± 0.00
Zeaxanthin	Derivatives only	27.43 ± 3.35	26.02 ± 1.94	22.92 ± 0.51

Table 1. MAPE of the baseline machine learning models for each preprocessing. Performance is evaluated as MAPE (%). Realized on 5 different shuffles of the dataset with the best preprocessing for each

Pigment	Comparison	PLS mean	Ridge mean	SVR mean	N pairs	t statistic	P value	Significant	Best model
Chlorophyll a	PLS vs Ridge	31.74	27.70	-	20	3.40	0.003	Yes	Ridge
Chlorophyll a	PLS vs SVR	31.74	-	30.58	20	0.96	0.350	No	No sig. diff.
Chlorophyll a	Ridge vs SVR	-	27.70	30.58	20	3.90	0.001	Yes	Ridge
Chlorophyll b	PLS vs Ridge	24.88	23.06	-	20	2.86	0.010	Yes	Ridge
Chlorophyll b	PLS vs SVR	24.88	-	20.76	20	5.48	0.000	Yes	SVR
Chlorophyll b	Ridge vs SVR	-	23.06	20.76	20	4.44	0.000	Yes	SVR
Lutein	PLS vs Ridge	19.67	19.20	-	20	0.57	0.575	No	No sig. diff.
Lutein	PLS vs SVR	19.67	-	16.91	20	2.93	0.009	Yes	SVR
Lutein	Ridge vs SVR	-	19.20	16.91	20	3.64	0.002	Yes	SVR
Violaxanthin	PLS vs Ridge	42.33	38.07	-	20	2.45	0.024	Yes	Ridge
Violaxanthin	PLS vs SVR	42.33	-	22.81	20	5.33	0.000	Yes	SVR
Violaxanthin	Ridge vs SVR	-	38.07	22.81	20	5.35	0.000	Yes	SVR
Zeaxanthin	PLS vs Ridge	32.18	30.98	-	20	1.86	0.078	No	No sig. diff.
Zeaxanthin	PLS vs SVR	32.18	-	18.67	20	9.95	0.000	Yes	SVR
Zeaxanthin	Ridge vs SVR	-	30.98	18.67	20	9.54	0.000	Yes	SVR

Table 2. Detailed statistical comparison of the baseline machine learning models. Performance is evaluated as MAPE (%). Realized on 5 different shuffles of the dataset averaged over the 4 preprocessing methods

Detailed parametrization

Table 3 provides additional details on the optimization procedure.

Parameter Category	Parameter Name	Value / Search Range
Fixed Training Settings	Batch Size	32 – 256
	Max Epochs	4,500 (with Early Stopping)
	Optimizer	Adam
	Learning Rate Scheduler	ReduceLROnPlateau (Factor=0.5, Patience=100)
	Early Stopping	Patience=100 epochs (Restore Best Weights)
	Loss Function (General)	Mean Absolute Percentage Error (MAPE, $\epsilon = 10^{-7}$)
	Loss Function (Zeaxanthin)	Log-Cosh Loss
Optuna Search Space	Learning Rate (Base)	8×10^{-4} – 3×10^{-3}
(Architecture)	Conv Layer 1 Filters	15 – 30
	Conv Layer 1 Kernel Size	15 – 25
	Conv Layer 2 Filters	20 – 35
	Conv Layer 2 Kernel Size	4 – 12
	Dense Layer Units	600 – 1,000
(Regularization)	Dropout Rate	0.08 – 0.20
	L2 Regularization (β)	0.05 – 0.15

Table 3. Details of the optimization procedure led with Optuna