

Human perception-based deep learning classification: an exploratory application to a large chronobiology dataset

Greta Guglielmetti ^{1,2,*}, Mario Castelán³, Diana Karen Miguel Sánchez³, Carlos Acuña⁴, David Martin¹,
Stephan Baumgartner^{1,2,5}, Alexander L. Tournier⁵

¹Institute of Integrative Medicine, University of Witten/Herdecke, Herdecke 58313, Germany

²Hiscia Research Institute, Society for Cancer Research, Arlesheim 4144, Switzerland

³Robotics and Advanced Manufacturing, Centre for Research and Advanced Studies of the National Polytechnic Institute, RamosArizpe 25900, Mexico

⁴School of Engineering and Sciences, Tecnológico de Monterrey, Monterrey 64849, Mexico

⁵Institute of Complementary and Integrative Medicine, University of Bern, Bern 3010, Switzerland

*Corresponding author. Institute of Integrative Medicine, University of Witten/Herdecke, Herdecke 58313, Germany. E-mail: g.guglielmetti@vfk.ch

Abstract

The Gold Fyfe dataset is the largest collection of daily chromatograms of one plant ($n = 29\,979$, covering over 27 years). Despite its high chronobiological relevance, it has so far been only partially explored. In this study, we tested a new characterization method of the chromatograms, based on qualitative features perceptible to human observers (i.e. gestalt recognition through kinaesthetic engagement) using deep learning. A trained human evaluator identified three gestalts and labelled chromatograms accordingly. Three deep learning models (one learning-from-scratch and two transfer learning approaches) were developed and evaluated across three data partitions for training validity, performance, and reliability. An explainability analysis was performed on one model, and the models' descriptive ability was compared with traditional image analysis descriptors previously used. A preliminary chronobiological application was tested on the dataset (PSDs and cross-correlation analysis). Evaluator consistency was 100%. The transfer learning models outperformed the learning-from-scratch model across all evaluation phases. The explainability analysis confirmed that the models relied on the same morphological features identified by the evaluator. Few correlations with traditional descriptors were found, suggesting an expansion of the descriptive palette. The preliminary chronobiological analysis revealed a correlation between one class and seasonal weather parameters. The proposed deep learning models represent a first attempt to achieve a qualitative characterization of metabolomic fingerprinting images guided by human perception. The models successfully extended the descriptive palette of the Gold Fyfe dataset with human-relevant descriptors, and their preliminary chronobiological application yielded promising results, warranting further investigation.

Keywords: plant chronobiology, image analysis, deep learning, human perception, metabolomic fingerprinting based on pattern formation

1. Introduction

The Gold Fyfe dataset is currently the largest chronobiology dataset of daily records available for one plant (Guglielmetti et al. 2024a). It consists of 29 979 capillary dynamolysis chromatograms collected over 27 years (Guglielmetti et al. 2024b). The plant under investigation is *Viscum album* L., a medicinal plant commonly known as mistletoe, of great interest for its

use in integrative cancer treatment (Büssing 2000, Maldacker 2006, Loef and Walach 2020, Ostermann et al. 2020). The scientific relevance of the dataset was demonstrated in our previous publication, where, after providing the first characterization of the dataset, the presence of infradian rhythms revealed its potential for chronobiological investigations (Guglielmetti et al. 2024a). However, the Gold Fyfe dataset to date remains

largely unexplored, highlighting the need for further investigation.

A major challenge for further evaluation of the dataset is the complexity and the uniqueness of the patterns displayed in Gold Fyfe’s chromatograms (see Fig. 1 and our previous publication Guglielmetti et al. (2024a)). No standardized technique exists for analysing the chromatograms generated by the specific capillary dynamolysis technique used in this dataset. Standardized evaluation criteria based on defined morphological features are available only for a variant of the method (i.e. the so-called WALA method) (Zalecka et al. 2010). However, this technique produces patterns that differ from those of Gold Fyfe chromatograms and is therefore not suitable for evaluating the Gold Fyfe dataset. Moreover, the proposed assessment (Zalecka et al. 2010) is based on a consensus-based panel of trained people. In this context, the size of the Gold Fyfe dataset renders this approach unworkable in practice, as involving multiple evaluators would be not only highly time-consuming and costly, but, more importantly, the evaluation of the whole dataset, i.e. close to 30K chromatograms, is a task beyond the capacity of human analysis. In this context, replacing human work with computers has the potential to ensure fast, cheap, reproducible, and consistent evaluation of the dataset. In our previous work, we developed a computer-based analysis in which we selected suitable features for the characterization of the chromatograms that were amenable to traditional image analysis tools (i.e. shape-related, colour, and textural features), defined, and explored elsewhere (Guglielmetti et al. 2024a). Although this analysis provided a good overall description of the analysed data, it did not integrate qualitative features identifiable through human vision.

Capillary dynamics analysis belongs to a larger class of methodologies that can be grouped under the umbrella of ‘metabolomic fingerprinting based on pattern formation’, hereafter referred to as metabolomic fingerprinting. Pattern recognition by human vision, as part of human perception, is traditionally an essential evaluation element of all these metabolomic fingerprinting methodologies. For copper chloride crystallisation, the most established among these metabolomic fingerprinting methods, an evaluation technique entirely based on human perception was developed (Doesburg et al. 2021). It is a consensus-based technique performed by a panel of trained evaluators built on two key pillars. The first one is that the description of the patterns is not only based on morphological characteristics but extended towards ‘the perception of salient coherent meaningful wholes or gestalts’ (Doesburg et al. 2021). The second one is that the evaluation of the patterns also includes the description of inner perceptions of the trained panellists arising as a response to pattern observation. The authors describe this process as kinaesthetic engagement since it ‘involves an embodied simulation of the growth, curvature and tension of the tree-like branches of the crystallisation images’. Evaluating crystallisation patterns using gestalt recognition through kinaesthetic engagement enhanced the performance of the panel in ranking tasks, suggesting the value and suitability of this technique in metabolomic fingerprinting evaluation (Doesburg et al. 2021).

Driven by these encouraging results, we aimed to extend the characterization of the dataset towards qualitative

features using the same core principle as proposed by Doesburg et al., i.e. gestalt recognition through kinaesthetic engagement. However, the scale of the task (29 979 chromatograms) requires computational assistance. A promising recent approach that can replicate various aspects of human perception (e.g. emotion recognition and aesthetic judgment) and has the advantage of scalability, is deep learning (Bodini 2019, Zhao et al. 2021). Therefore, in this preliminary work, we explored the feasibility of using this technique to characterize the Gold Fyfe dataset, according to the human ability of gestalt recognition through kinaesthetic engagement.

For this purpose, three main classes of patterns were identified by a trained evaluator as representative of three gestalts that can be perceived in the chromatograms: Uprightness, Lightness, and Heaviness. Supervised deep learning models were then developed. They were first evaluated across different partitions of the data to assess the validity, performance, and reliability of their training. Next, the image regions that drove the decisions of the best-performing one were assessed using an explainability analysis. Finally, their capacity to extend the dataset's current descriptive palette was assessed. For this, their outputs across the entire Gold Fyfe dataset were compared with those of the traditional image analysis descriptors previously used to characterize this dataset. Finally, as a preliminary application, the models were used to assess their potential value for chronobiological studies.

2. Materials and methods

2.1 The Gold Fyfe dataset

From 1st November 1958 to 14th October 1985 daily samples of *V. album* L. grown on apple trees (*Malus domestica* Borkh.) were collected and analysed at the Hiscia Research Institute (Society for Cancer Research, Arlesheim, Switzerland). With a total number of 27 979 chromatograms, this dataset is the largest of its kind (Guglielmetti et al. 2024a, 2024b). The dataset comprises two subgroups of chromatograms that differ in the sample concentration. In the present study, as in our previous work, we selected these chromatograms made with the higher (50%) sample concentration ($n=19\,109$). This is because they are visually more diverse and informative than the lower-concentration chromatograms.

The fingerprints were made using a technique called capillary dynamolysis, which produces coloured patterns on chromatography paper (Kolisko 1953, Steffen 1983, Skjerbaek et al. 2004, Zalecka et al. 2010, Kokornaczyk et al. 2021). In the present case, the chromatograms were made with AuCl_3 as a reagent, which typically produces pink, purple, and yellow patterns (see fig 1 of Guglielmetti et al. (2024a)). For further specifications about the technique and sample preparation, we refer to earlier publications (Fyfe 1975, Guglielmetti et al. 2024a).

The chromatograms were scanned between June 2020 and April 2021 in order to analyse them with computer-based image analysis methods (Guglielmetti et al. 2024a).

2.2 Classes description

A human evaluator (GG) studied the complete dataset over the course of a year during the scanning process. Each

Ideal and Optimal Partitions were univocally labelled (i.e. one image assigned to one class), whereas the images of the Suboptimal Partition received non-univocal labels (i.e. one image could be assigned to more than one class).

2.3.1 The ideal partition

Images were selected according to two criteria: the evaluator's confidence in assigning each image to a specific class, and the clarity with which the image expressed the class's gestalt. Only images considered by the evaluator truly representative of their category were included. Thus, a balanced partition of 245 univocally labelled images per class (735 images in total, 3.8% of the dataset) was created.

2.3.2 The optimal partition

Of the 1000 randomly selected images, only those that received univocal labels were included in the Optimal Partition. This partition was referred to as ‘Optimal’ because it sat between the Ideal Partition (hand-picked, highly representative patterns) and the Suboptimal Partition (random selection, coexisting gestalts requiring non-univocal labels). Unlike the Ideal Partition, these images were not considered representative of their class by the evaluator, as they did not fully carry one of the three gestalts. This process yielded a partition of 536, 80, and 54 univocally labelled images for the Uprightness, Heaviness, and Lightness classes, respectively (670 images in total, 67% of the random selection, 3.5% of the dataset). The class imbalance reflected the dataset composition.

2.3.3 The suboptimal partition

Of the 1000 randomly selected images, only those with non-univocal labels were included in this partition. Each image was assigned a score for each of the three classes: 0 indicated no visible characteristics, 1 indicated some, and 2 indicated predominant characteristics. Three non-univocal label types were possible: the pattern equally showed characteristics of two classes (1;1;0 and permutations); the pattern showed characteristics of two classes but one more pronounced (2;1;0 and permutations); the pattern did not present characteristics of any of the classes (0;0;0). This process yielded a partition of 32, 219, and 79 images, accordingly labelled images (330 images in total, 33% of the computer selection, 1.7% of the dataset). The need for non-univocal labels was symptomatic of the dataset, which demonstrated its inherent fuzziness, as the transition from one gestalt to another could be observed across a series of chromatograms in chronological order. A more detailed description of the classes' composition in each group of images is in the [Supplementary Material](#).

2.4 Evaluator consistency

The consistency of the evaluator's labelling was tested on the Ideal Partition, as it was the partition chosen for network training. Since the gestalt recognition through kinaesthetic engagement relies on the evaluator training and the consistency test was run long after the creation of the partition (>1 year), GG was retrained for 1 week on the Ideal Partition. The training consisted of studying the three classes again by kinaesthetically engaging with the corresponding images. One week after

the training, 20% of the images ($n=49$ images per class, $n=147$ images in total) were randomly selected. AT assigned blind codes to the images so that GG could not identify them from the original labelling session, and then GG re-labelled them independently.

2.5 Deep learning models

Three models were developed: one with a learning-from-scratch approach and two with transfer learning.

The learning-from-scratch model (herein referred to as the LFS Model) was conceived as an exploratory approach to test the applicability of deep learning-based image analysis on the Gold Fyfe dataset. It was designed to assess whether relevant visual primitives and discriminative features could be learned directly from the dataset itself, without relying on pre-trained weights. This was considered feasible because the training set was constructed from highly representative and relatively unambiguous examples. Full architectural specifications are provided in the [Supplementary Material](#).

The transfer learning models were developed to address potential limitations of the LFS Model, (e.g. the risk of overfitting and the associated reduction in performance on unseen examples). One implementation utilized DenseNet121 (herein referred to as the DTL Model) (Huang et al. 2017) and the other one used ResNet50 (herein referred to as the RTL Model) (Koonce 2021). Both networks were initialized with ImageNet pretrained weights (Deng et al. 2009) and subsequently fine-tuned. Full architectural specifications are provided in the [Supplementary Material](#).

2.6 Traditional descriptors

Descriptors computed with traditional image analysis (hereafter referred to as Traditional Descriptors) relating to the outline, colour, and texture of the images were also used to characterize the images of the Gold Fyfe dataset. The details of the extraction of the Traditional Descriptors have been reported previously (Guglielmetti et al. 2024a). In brief, traditional image analysis was performed in batch on the whole dataset using Fiji ImageJ (Guglielmetti et al. 2024a). In total, 12 descriptors were calculated: Area% St-Frame, Perimeter, Area% Bounding Box, Height, Prominence, Entropy, Ang_2nd_Moment, Inv_Diff_Moment, Purple Area, Yellow Area%, Yellow_Y, Yellow_X. Yellow_X was used as a negative control. It is the horizontal component of the centre of mass of the yellow area of the patterns, which is randomly distributed around the centre of the chromatography paper.

2.7 Weather data

Weather parameters measured at the meteorological station of Basel-Binningen during the same time span as our dataset (1958–1985) were obtained from MeteoSwiss (Switzerland). A selection of 12 representative parameters was chosen: temperature, wind, fresh snow, cloud cover, atmospheric pressure, vapour pressure, precipitation, sunshine duration, air humidity, days with rain and snow, days with graupel, days with thunderstorm (tre200d0, fkl010d0, hns000d0, nto008d0, prestad0, pva200d0, rhs150d0, sre000d0, ure200d0, w2p001d0, w3p001d0, w4pclod0 parameters from MeteoSwiss).

2.8 Data analysis

2.8.1 Evaluation of the deep learning models on data partitions

The models were evaluated on the data partitions using a three-phase pipeline to assess: training validity (Phase 1), training performance (Phase 2), and training reliability (Phase 3). For Phase 1, 10 independent training runs were performed for each model to reduce the influence of stochastic factors such as weight initialization and data shuffling; for Phases 2 and 3, the models were run only once. The training was always done on the Ideal Partition, while the partition used for testing differed between the three phases. More specifically:

- Phase 1—Training validity: training on 80% of the Ideal Partition ($n=588$ images in total), testing on 20% of the Ideal Partition ($n=147$ images in total).
- Phase 2—Training performance: training on 100% of the Ideal Partition ($n=735$ images), testing on 100% of the Optimal Partition.
- Phase 3—Training reliability: same training as Phase 2 (i.e. 100% of the Ideal Partition), testing on 100% of the Suboptimal Partition.

To evaluate the models' ability to assign each image to a class, the likelihood or confidence scores produced by the models' final layers were used. The scores, processed through a SoftMax activation function (Goodfellow et al. 2016), indicated the probability that a given image belonged to a particular class. These probability scores were normalized to sum to 1. Every image was scored using this procedure. The images were assigned to the class with the highest score. To evaluate the models, several parameters were calculated. Accuracy (i.e. the proportion of correctly classified samples out of the total number of samples), precision (i.e. the percentage of images predicted as belonging to a class that actually belong to that class), recall (i.e. the percentage of images that truly belong to a class that were correctly identified as belonging to that class), and F1 (i.e. the harmonic mean of precision and recall, providing a single balanced measure of the model's performance). Such parameters were reported both class-wise and as macro, meaning they were computed as an unweighted average of the per-class scores. Additionally, $P_{\max}$, defined as the maximum value in the SoftMax output vector, was calculated. It was used as a measure of prediction confidence, with higher values indicating greater certainty in the assigned class. In Phase 3, only $P_{\max}$ was calculated, as the non-univocal labels of the Suboptimal Partition prevented the calculation of the other parameters.

2.8.2 Deep learning explainability analysis

To examine the image regions that contributed to the model's decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to the best-performing model according to the metric results (Selvaraju et al. 2020). For this analysis, 8 images were randomly selected for each confidence level determined in Phase 2 per class (24 images per class, 72 images in total).

2.8.3 Cross-correlation with traditional descriptors

The models were applied to the entire dataset ($n=19\,109$ images). Rather than using the probability scores produced by the models' last layer to classify individual images, the scores themselves were extracted for all three classes across the dataset. When multiple images were available for a given day, their scores were averaged, yielding a single mean score per class per day. This produced scores for 9627 days; the remaining 218 days (2.21% of the total time span) were filled using linear interpolation, yielding a total of 9845 days per class per model. For each model, Pearson cross-correlations were calculated between these scores and the Traditional Descriptors.

2.8.4 Preliminary chronobiological application

As a preliminary exploratory application of the newly characterized dataset to chronobiology, the same daily class scores calculated for the cross-correlation with the Traditional Descriptors were then analysed as a time series. Welch's method was applied to compute the power spectral density (PSD) for each class and model, revealing underlying periodicities, while Pearson cross-correlations were calculated against the weather parameters. The data were processed and statistically analysed using Python 3.10 (Van Rossum and Drake 2009) on a standard PC (Latitude 3440, Dell, Round Rock, USA).

3. Results

3.1 Evaluator consistency

In the evaluator consistency test, the evaluator achieved a score of 100%, i.e. all images ($n=49$ per class, 147 in total) were correctly assigned to the corresponding class.

3.2 Evaluation of the deep learning models

3.2.1 Phase 1—training validity

Table 1 shows the results of the first evaluation phase. Overall, the LFS Model achieved lower performance and greater variability across the 10 independent training runs than both transfer learning models, which performed comparably (accuracy: ~84% for LFS vs. ~89% for DTL and RTL). At the class level, Lightness and Uprightness were the best-performing classes across all three models, while Heaviness consistently showed the lowest scores. The relatively small standard deviations across runs suggested that the training process was stable and that the models captured robust features of the chromatographic structures.

3.2.2 Phase 2—training performance

Phase 2 was the first phase in which training and testing used distinct partitions, making it appropriate to assess whether data augmentation would improve generalization. Accuracy with data augmentation was consistently lower than without it across all three models (Table 2), possibly because the augmented transformations introduced visual distortions unrepresentative of the natural variability in the chromatograms. Data augmentation was therefore discarded, and all subsequent results refer to models trained without it.

Table 2 shows the results of the second evaluation phase. The LFS Model showed a substantial drop in performance

Table 1 Performance of the models in Phase 1 (training validity).

Model	Accuracy	Precision		Recall		F1	
		Macro	Class-wise	Macro	Class-wise	Macro	Class-wise
LFS	83.7 ± 4.3	85.2 ± 3.1	U: 86.5 ± 6.4 H: 79.7 ± 11.1 L: 89.2 ± 7.3	83.7 ± 3.9	U: 83.5 ± 11.3 H: 82.5 ± 11.4 L: 85.1 ± 6.9	83.6 ± 4.0	U: 84.3 ± 5.7 H: 79.9 ± 5.9 L: 86.7 ± 3.3
DTL	89.0 ± 3.8	89.3 ± 3.8	U: 87.8 ± 6.1 H: 85.0 ± 6.4 L: 95.0 ± 2.3	89.0 ± 3.9	U: 90.6 ± 5.5 H: 84.9 ± 6.5 L: 91.2 ± 4.5	88.9 ± 3.9	U: 89.1 ± 4.9 H: 84.7 ± 4.5 L: 93.0 ± 2.0
RTL	88.6 ± 2.8	89.0 ± 2.8	U: 91.1 ± 4.7 H: 83.9 ± 5.4 L: 91.7 ± 3.9	88.6 ± 2.8	U: 91.2 ± 6.0 H: 84.7 ± 5.3 L: 90.0 ± 4.4	88.6 ± 2.8	U: 91.0 ± 3.9 H: 84.2 ± 4.1 L: 90.7 ± 2.7

Values are reported as percentages, with mean ± standard deviation across 10 independent training runs. Metrics include overall accuracy and precision, recall, and F1-score, both macro and class-wise, for the three classes. U, uprightness; H, heaviness; L, lightness.

Table 2 Performance of the models in Phase 2 (training performance).

Model	Accuracy	Precision		Recall		F1	
		Macro	Class-wise	Macro	Class-wise	Macro	Class-wise
LFS	52.1	48.1	U: 97.3 H: 29.9 L: 17.0	62.6	U: 47.4 H: 73.8 L: 66.7	44.4	U: 63.7 H: 42.6 L: 27.1
DTL	79.4	59.0	U: 96.2 H: 38.7 L: 42.1	69.8	U: 84.1 H: 51.3 L: 74.1	62.5	U: 89.8 H: 44.1 L: 53.7
RTL	79.4	62.2	U: 99.1 H: 40.8 L: 46.8	78.1	U: 80.2 H: 72.5 L: 81.5	69.7	U: 88.5 H: 52.3 L: 59.5

Values are reported as percentages. Metrics include overall accuracy and precision, recall, and F1-score, both macro and class-wise, for the three classes. U, uprightness; H, heaviness; L, lightness.

compared with the two transfer learning models (accuracy: 52.1% vs. 79.4% for both DTL and RTL). At the class level, Uprightness clearly outperformed the other two classes across most metrics. The performance of Lightness and Heaviness, however, reversed between the LFS Model and the transfer learning models: Heaviness scored higher than Lightness for LFS, while the opposite held for DTL and RTL. Notably, RTL consistently achieved higher recall than DTL for the minority classes (Heaviness and Lightness), suggesting greater sensitivity to subtle chromatographic variations.

The distribution of $P_{\max}$ for correct and incorrect predictions was also examined. For all models, correct predictions were associated with higher mean $P_{\max}$ values than incorrect ones, supporting the use of $P_{\max}$ as a practical indicator of prediction reliability. Three confidence levels were therefore defined (low, medium, and high; see [Supplementary Material](#)), and the results are summarized in [Table 3](#). The LFS Model assigned the majority of images to the high-confidence level, yet achieved only modest accuracy there, indicating that it was often wrong. In contrast, both transfer learning models showed accuracy that rose markedly with confidence level, with RTL reaching the highest accuracy among high-confidence predictions.

3.2.3 Phase 3—training reliability

As mentioned above, due to the non-univocal labels of the images in the Suboptimal Partition, this evaluation phase was restricted to analysing $P_{\max}$ values. Results are provided in [Table 4](#). The two transfer learning models showed a clear shift in predictions towards the low-confidence level, appropriately mirroring the partition's ambiguity. The LFS Model, in contrast, skewed towards high confidence, with the majority of images assigned to the high-confidence level. This pattern reinforced the tendency towards overconfident classification already observed in Phase 2, suggesting that the LFS Model produced unreliable confidence estimates under conditions of morphological ambiguity.

3.2.4 Deep learning explainability analysis

The explainability analysis, using Grad-CAM maps, was conducted on RTL because it showed consistently strong performance across the three evaluation phases and outperformed the other models, particularly in Phase 2. For Uprightness, activations formed a horizontal band across the central region of the pattern, corresponding to the vertical columnar structures characteristic of this class ([Supplementary Figure S1](#)). For Heaviness, activations concentrated in the lower-middle

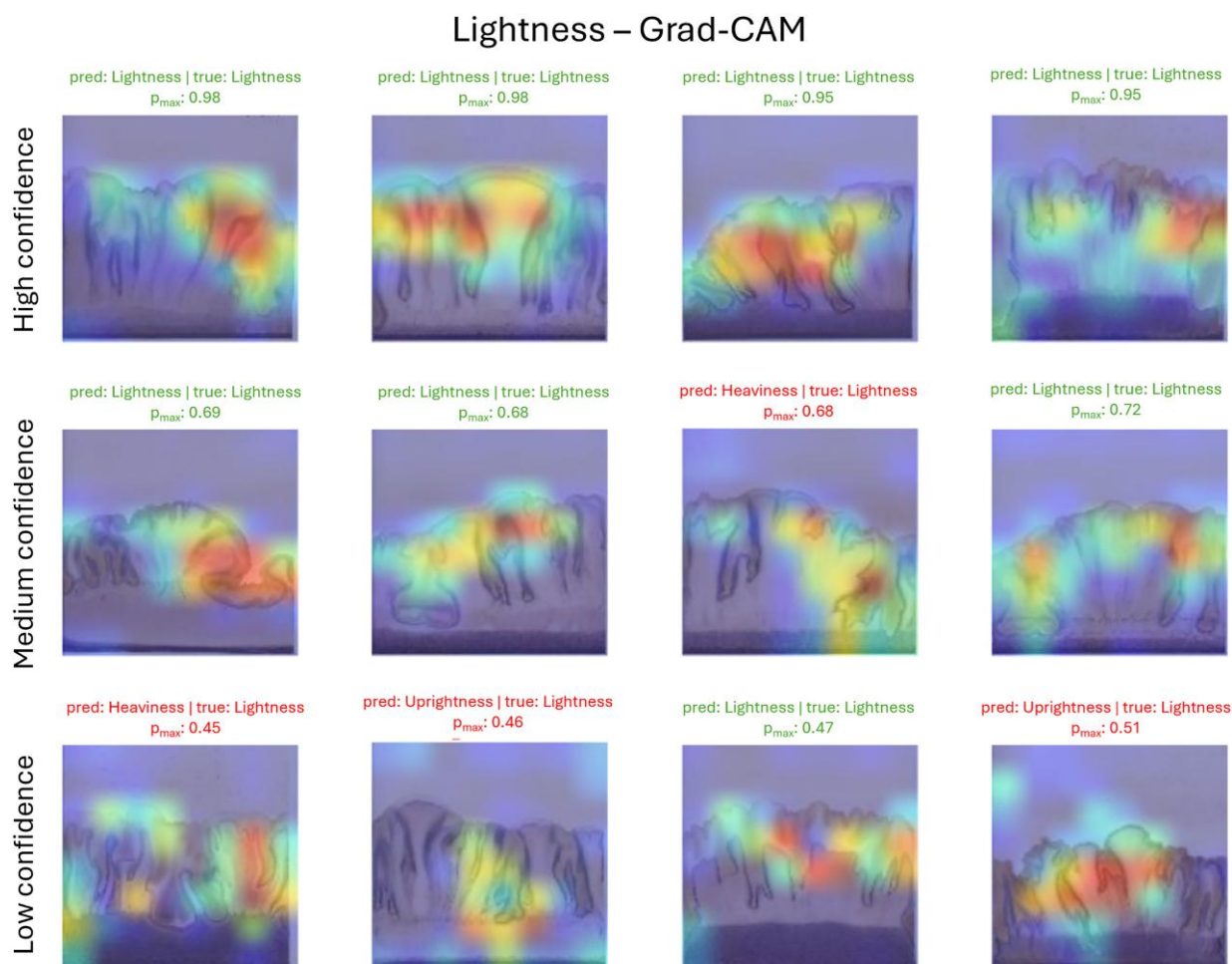


Figure 2 Representative examples of Grad-CAM activation map for the Lightness class (full set in [Supplementary Material](#)). The activation map is superimposed on the original image using a jet colour scale, where red indicates regions of highest activation intensity and blue indicates lower intensity. High-intensity regions contributed more strongly to the model's classification decision. The predicted class (pred), true class (true), and P_{max} value are reported beneath each image, displayed in green when the prediction matches the true label and in red when it does not.

main peaks appearing across the three models: ~ 360 days, ~ 200 days, and ~ 120 days.

The peak around 360 days (about 1 year) was consistently present across the classes/models. The highest PSD were reached by the Uprightness and Lightness classes. While Lightness was the Highest peak for the LFS and DTL models (~ 11 , ~ 8 , respectively), followed by Uprightness (~ 9 , ~ 6 , respectively), in the RTL Model, the two were swapped (~ 10 , ~ 7 for Uprightness and Lightness, respectively). For Heaviness, the peak power decreased across the models from ~ 7 in the LFS Model, to ~ 2 in the DTL Model and to ~ 1.5 in the RTL Model.

The peak around 200 days (about a 6-month period) was clearly present across the models only for Uprightness and Lightness classes. Uprightness peaks showed the highest PSD in the LFS Model (~ 7), a milder peak (~ 4) in the RTL Model and a weaker one (~ 2.5) in the DTL Model. The Lightness peaks showed a similar power pattern across the three models, although with generally milder intensity (~ 3 , ~ 2 , ~ 1.5 for LFS, RTL, and DTL, respectively). Heaviness presented a peak around that region but shifted slightly around 170/180 days. This may reflect a greater sensitivity of the LFS Model

to dominant temporal patterns in the dataset, leading to stronger periodic components in the PSD spectra. However, this did not translate into better predictive performance, suggesting that the model may have amplified recurrent patterns that were highly specific to the training data rather than broadly generalizable.

The peak around 120 days (about a 4-month period) was stably present across the classes/models. For Uprightness and Lightness, the peaks showed consistent PSD of around 2.5 and 1.5 across models, respectively. For Heaviness, the peak PSD decreased across the models from ~ 1.5 in the LFS Model, to ~ 0.6 in the DTL Model and to ~ 0.4 in the RTL Model.

In summary, the three models showed fairly consistent spectral behaviour across the three main periodicities. Uprightness and Lightness consistently produced the highest PSD peaks, while Heaviness was the weakest spectral component across all periods and models. Among the three models, the LFS Model generally exhibited the highest peak PSD.

3.3.2 Cross-correlation with weather parameters

The identification of a dominant ~ 360 -day peak in the PSD spectra, together with lower-intensity peaks at ~ 120 and

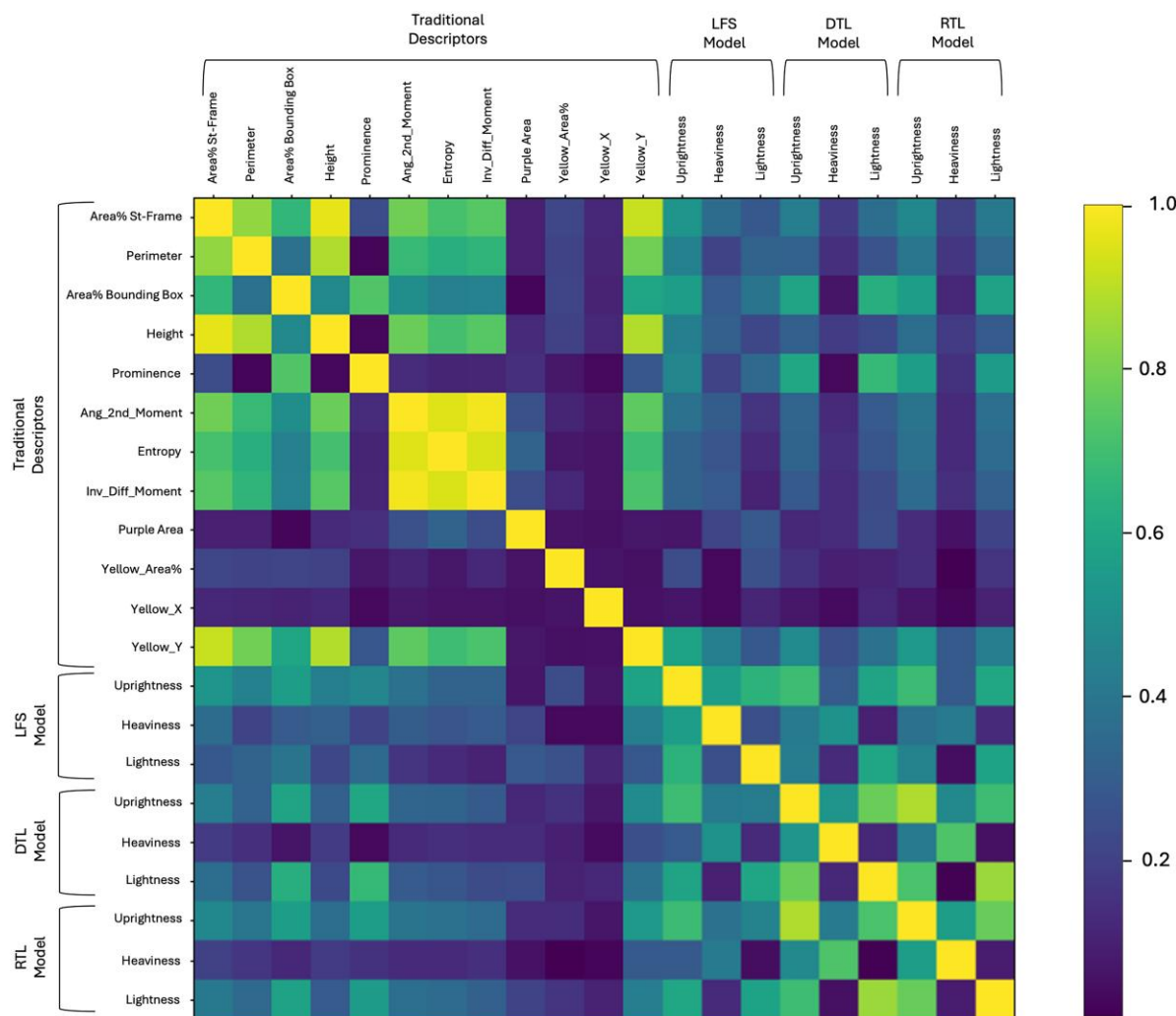


Figure 3 Cross-correlation matrix between traditional descriptors and deep learning models LFS, DTL, and RTL. The coloured bar on the right provides grades of colours as a reference, ranging from highest (1 = yellow) to no correlation (0 = deep blue).

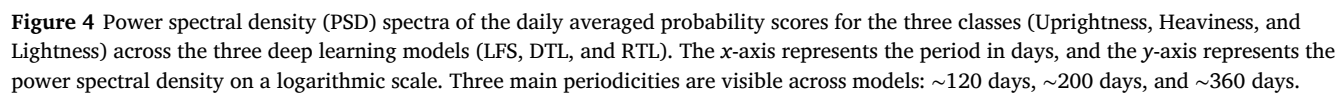
~200 days (i.e. potentially harmonics of the 360-day peak), suggested that the models' probability scores may reflect seasonal environmental variation. To explore this, we calculated correlations between a set of meteorological parameters already used in our previous publication (Guglielmetti et al. 2024a) and the time series data obtained from the deep learning models (i.e. LFS, DTL, and RTL) and from the Traditional Descriptors. Figure 5 shows this cross-correlation; complete data are available in Supplementary Table S5.

The correlations already found in our previous publication between the Traditional Descriptor Purple Area and temperature and vapour pressure were the highest in the matrix (-0.4 for both weather parameters). However, the only other visible correlations (i.e. ≥ 0.2) are between the same weather parameters (i.e. temperature and vapour pressure) and Lightness and Uprightness of the deep learning models. In particular, Lightness for the LFS Model showed the strongest anti-correlations (-0.3 for both weather parameters), followed by Lightness for the DTL and RTL models (DTL: -0.2 ; RTL: -0.1 , for both weather parameters). Additionally, Uprightness for the RTL Model showed a mild correlation with vapour pressure (0.2).

4. Discussion

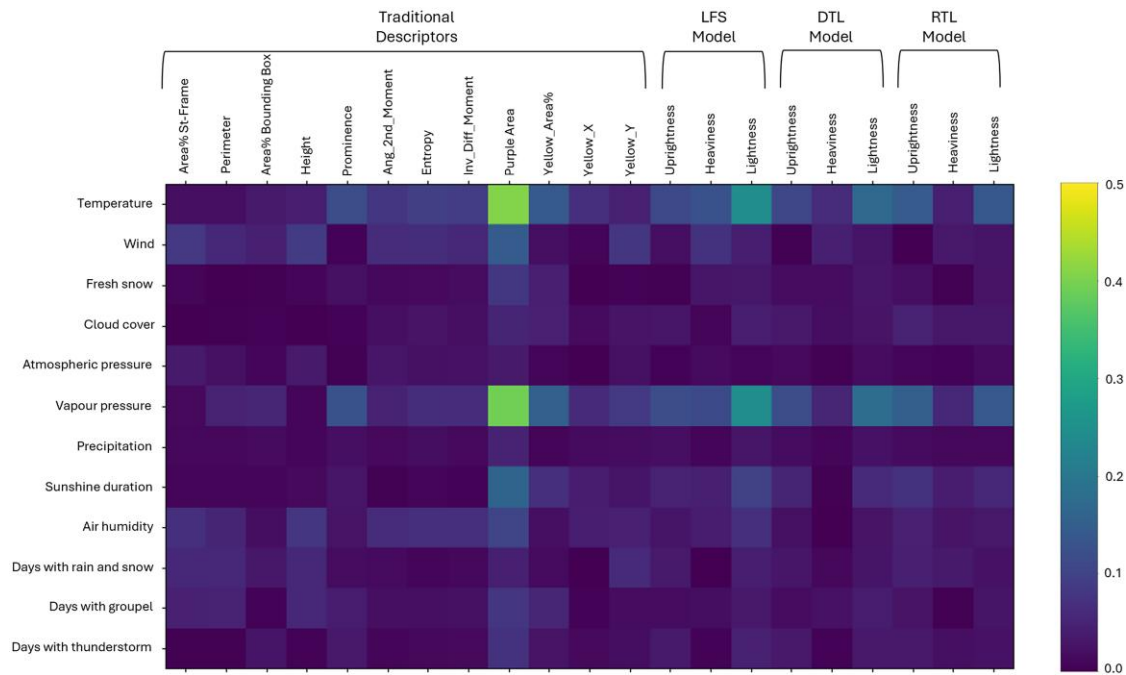
The Gold Fyfe dataset, previously characterized only through traditional image analysis (Arnheim 1954), has remained largely unexplored. This study aimed to expand the descriptive palette of the dataset with human-relatable descriptors by combining human gestalt recognition through kinaesthetic engagement with deep learning, ensuring scalability to the full dataset. To this end, a trained human evaluator identified three classes of patterns in the images, corresponding to three gestalts: Uprightness, Heaviness, and Lightness. Three deep learning models, one with a learning-from-scratch approach (i.e. LFS Model) and two based on transfer learning (i.e. DTL and RTL models), were developed and tested on three data partitions with either univocal or non-univocal labels according to the representativeness and clearness of the patterns.

The results of the three evaluation phases showed that the transfer learning models performed consistently better than the LFS Model, which was revealed to be overconfident in its predictions. Of the two transfer learning models, the RTL Model achieved higher recall for the two similar classes, Lightness and Heaviness, suggesting greater sensitivity to



The results of the cross-correlation analysis between the deep learning models and the Traditional Descriptors suggested an expansion of the descriptive palette for the Gold Fyfe dataset. The classes of the models were mutually anti-correlated, indicating they are not redundant and describe different features.

The preliminary application of the models for chronobiological purposes on the Gold Fyre showed interesting results. The PSD analysis revealed three main frequencies across models and classes at around 120, 200, and 360 days, particularly for the LFS Model. The dominant periodicity was around 360 days, suggesting an annual periodicity, of which the other two peaks may be harmonics. The cross-correlation with weather parameters, especially with the LFS Model, supported this interpretation, although the correlation was modest.



are present in the dataset. It should be noted that the three investigated classes were not intended to provide a comprehensive, gestalt-based characterization of the dataset. They were selected because of their recurrence in the dataset, not because they were considered the only ones present.

Finally, the chronobiological analysis raised an interesting point: the LFS Model, despite underperforming the transfer learning models in classification accuracy and showing overconfident predictions, produced the highest PSD peak and the strongest correlations with weather parameters. One possible explanation is that the LFS Model, unconstrained by pre-learned feature representations, captured broader structural variations in the chromatograms that co-vary more strongly with seasonal changes but are less useful for class discrimination. This dissociation between classification performance and chronobiological sensitivity warrants further investigation, as it suggests that the two tasks may benefit from different model properties. That said, the primary contribution of this work lies in the development of a gestalt-based classification mediated by deep learning, an approach applied here for the first time, which has been shown to successfully extend the descriptive palette of the Gold Fyfe dataset. The chronobiological application, while promising, remains preliminary and will require further refinement to fully exploit the potential of the proposed descriptors.

The complexity of the Gold Fyfe dataset, together with the large number of currently unlabelled images, suggests that future work may benefit from unsupervised learning approaches capable of identifying recurrent morphological structures without relying on extensive manual annotation. The present supervised approach provides an important foundation, as it establishes an initial set of meaningful descriptors, interpretable classes, and model behaviours against which future unsupervised patterns may be compared.

5. Conclusion

The deep learning models developed in this study represent the first attempt to achieve, by computational means, a qualitative characterization of metabolomic fingerprinting images guided by human perception, specifically, gestalt recognition through kinaesthetic engagement. The models were shown to successfully extend the descriptive palette of the Gold Fyfe dataset with new, human-relatable descriptors. Their preliminary application to chronobiology yielded promising initial results that warrant further investigation.

Author contributions

Writing—original draft: G.G.; writing—review and editing: G.G., A.L.T., M.C., S.B., D.M.; conceptualization: G.G., S.B., M.C.; methodology: G.G., S.B., M.C.; investigation: G.G., M.C., C.A., D.K.M.S.; formal analysis: G.G., A.L.T.; data curation: G.G.; visualization: G.G.; supervision: A.L.T., S.B., D.M., M.C.

Supplementary material

Supplementary material is available at *in silico Plants* online.

Conflicts of interest

None declared.

Funding

The authors received no funding for this work.

Data availability

The image dataset used in this study was presented in a previous paper (Arnheim 1954) and is available on request DOI: <https://doi.org/10.5281/zenodo.11032266>

References

- Arnheim R. *Art and Visual Perception: A Psychology of the Creative Eye*. Berkeley: University of California Press, 1954.
- Bodini M. Will the machine like your image? Automatic assessment of beauty in images with machine learning techniques. *Inventions* 2019;4:34. <https://doi.org/10.3390/inventions4030034>
- Büssing A. Biological and pharmacological properties of *Viscum album* L. In: Büssing A (ed.) *Mistletoe: the Genus Viscum*. Amsterdam, The Netherlands: Harwood Academic Publishers, 2000, 123–82.
- Deng J, Dong W, Socher R *et al*. Imagenet: a large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Miami, F: IEEE, 2009.
- Dictionary CE. *Definition of gestalt from the Collins English Dictionary*, 2020. <https://www.collinsdictionary.com/dictionary/english/gestalt> [cited 2026 May 4].
- Doesburg P, Fritz J, Athmann M *et al*. Kinesthetic engagement in gestalt evaluation outcores analytical ‘atomic feature’ evaluation in perceiving aging in crystallization images of agricultural products. *PLoS One* 2021;16:e0248124. <https://doi.org/10.1371/journal.pone.0248124>
- Doesburg P, Huber M, Andersen J-O *et al*. Standardization and performance of a visual gestalt evaluation of biocrystallization patterns reflecting ripening and decomposition processes in food samples. *Biological Agriculture and Horticulture* 2015;31:128–45. <https://doi.org/10.1080/01448765.2014.993705>
- Fyfe A. The signature of the planet mercury in plants—capillary dynamic studies. *Br Homeopath J* 1974;62:201–32. 63 (1), 26–60; 63 (2), 111–124. [https://doi.org/10.1016/S0007-0785\(73\)80032-8](https://doi.org/10.1016/S0007-0785(73)80032-8)
- Fyfe A. *Moon and Plant—Capillary Dynamic Studies*, 2nd edn. Arlesheim: Society for Cancer Research, 1975.
- Fyfe A. *Die Signatur der Venus im Pflanzenreich—Kapillardynamische Untersuchungsergebnisse*. ed. V.f.K. Forschungsinstitut Hiscia, Arlesheim, Schweiz. Stuttgart: Verlag Freies Geistesleben, 1978, 90.
- Galotti KM. *Cognitive Psychology in and out of the Laboratory*, 5 edn. CA: Thousand Oaks, 2014.
- Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA: MIT Press, 2016.

