

How Spectral Properties and Machine Learning Can Categorize Twin Species - Based on *Diachrysia* Genus

Krzysztof Dyba

Adam Mickiewicz University in Poznań

Roman Wąsala

University of Life Sciences in Poznań

Jan Piekarczyk

Adam Mickiewicz University in Poznań

Elżbieta Gabala

Institute of Plant Protection

Magdalena Gawlak

Institute of Plant Protection

Jarosław Jasiewicz

Adam Mickiewicz University in Poznań

Henryk Ratajkiewicz (✉ henryk.ratajkiewicz@up.poznan.pl)

University of Life Sciences in Poznań

Research Article

Keywords: machine learning, twin species, Noctuidae, absorption

Posted Date: April 7th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-390055/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Confirmation of distinctiveness or taxonomic affinity awaits new evidence in many twin species. In our work we used noninvasive point reflectance spectroscopy in the range from 400 to 2100 nm coupled with machine learning to study scales on the brown and golden iridescent areas on the dorsal side of the forewing of *Diachrysia chrysis* and *D. stenochrysis*. The basis for the study a statistically significant collection of 95 specimens gathered during 23 years in Poland. The numerical part of an experiment included two independent discriminant analyses: stochastic and deterministic. The more sensitive stochastic approach achieved average compliance with the species classification made by entomologists at the level of 99-100%. It demonstrated high stability against the different configurations of training and validation sets, hence strong predictors of *Diachrysia* siblings distinctiveness. Both methods resulted in the same small set of relevant features, where minimal fully discriminating subsets of wavelengths were three for glass scales on the golden area and four for the brown. The differences between species in scales primarily concern their major components and ultrastructure. In melanin-absent glass scales, this is mainly chitin configuration, while in melanin-present brown scales, melanin reveals as an additional factor.

Introduction

In the last decades, numerous attempts were made to revise the insect's taxonomy. We became interested in the Noctuidae family with about 25 000 known species¹ representing one of the most species-rich family of Lepidoptera. The noctuids subfamily Plusiinae (Boisduval, 1828) with *Plusia* (Ochsenheimer, 1816) as its type is distributed worldwide except in the Antarctic². This popular and easily recognizable noctuid group is, as one might expect, monophyletic with numerous morphological, bionomical and biogeographical evidence for their monophyly. However, the situation is so far not clear, despite the intense studies made on the Plusiinae in the last two decades^{2,3}. The taxonomic interpretation of twin species: *Diachrysia chrysis* (Linnaeus, 1758) and *Diachrysia stenochrysis* (Warren, 1913) (Fig. 1) is one of the biggest problems in this subfamily. In 1961 Kostrowicki distinguished *Plusia tutti* as a new species from *D. chrysis* on the basis of certain external and male genital features⁴. It turned out, however, that *P. tutti* and *D. stenochrysis* are the same species. This gave rise to a series of studies chronologically stimulated by Lempke⁵, Urbahn^{6,7} and Rezbanyai-Reser⁸. As a result of these studies which involved breeding and morphological examination of the genital organs, it was concluded that some individual moths could not be attributed to one or another species due to too large differences in genitalia, wing pattern and caterpillar coloration. Despite this, the taxonomic separateness of both species was questioned, which generated a number of further studies. With the use of specific sex pheromones by Priesner⁹, taxonomic research on *Diachrysia* spp. entered a new path. They were continued, among others by Bruun¹⁰, Svensson et al.¹¹, Löfstedt et al.¹² and Inomata et al.¹³, however, did not solve the problems in species classification. Soon, new insight into the taxonomy of these twin species was made possible by research on reproductive isolation of pheromone-trapped male *Diachrysia* using allozyme analysis initiated by Svensson et al.¹¹. Further researches were continued with the use of the cytochrome oxidase

subunit I^{14,15}. However, the problem of taxonomy within the *Diachrysia* genus remained unresolved, and Hausmann et al.¹⁶ indicated that further research is needed at a broad geographical scale.

The declining number of taxonomists experienced in morphological insect evaluation¹⁷ results in an increased use of automatic identification methods^{18,19}. Reflectance spectroscopy (RS) methods are complementary to expert or genetic classification²⁰. On the contrary to the visual evaluation, RS provides broader spectra including the ultraviolet, near and mid-infrared range. The advantage of RS is the potential to quickly obtain a large amount of information about the studied object during a single and possibly non-invasive measurement.

In recent years, RS has been applied in various disciplines of biology^{21,22}. Several papers in entomology involving visible (VIS) and near-infrared (NIR) spectroscopy applications coupled with numerical methods have been recently reported (see: Johnson and Naiker¹⁹ for a comprehensive list).

Many studies refer to spectral properties of butterflies and moths but most of them have considered only reflectance in human-visible wavelengths (380–780 nm) in which coloration is perceivable. The color of the wing can be pigment-dependent and a separate phenomenon is the structural colors of the wing, which was the subject of in-depth research conducted also on the Noctuidae family^{23–25}.

Coloration enables expert classification as well as can be suitable for automatic species identification. Kaya and Kayci²⁶ using neural networks trained on the RGB images and its texture created a classifier which distinguishes 14 species of butterflies with over 90% accuracy. Their collection covered only examples with a relatively different color pattern of wings. However, there exist many lepidopteran species which cause identification problems because wing coloration is not a diagnostic feature.

Few studies considered butterflies and moths wing reflectance properties beyond visible range of spectra^{22,27}. This region of the spectrum is useful in the detection of stored-grain insects^{28,29}. The NIR and SWIR (short-wave infrared) ranges comprise approximately 50% of total incident radiation³⁰. In this region of the spectrum, information on the structure, water and fat content, and the presence of chitin in the studied objects is encrypted^{31,32}. Nevertheless, only a few chemical compounds are identified, mainly bonds occurring in the structure. Gebru et al.³³ showed three premises that can be useful in the detection and identification of insects: low absorption of radiation by melanin over the entire SWIR range, strong absorption by water in the band 1470–1550 nm, and 1320 nm wavelength, which has been indicated as unaffected by both melanin and water. Combining information from SWIR and NIR was useful for insect species classification as well as physiological status, age and sex assessment^{33–35}. However, studies to date do not provide information on species determination in the Lepidoptera order based on the wing reflectance spectrum in the NIR and SWIR regions.

Detailed information about the optical properties and structure of butterfly wing scales is provided by microscopic studies. For many years, the subject of such research has been the phenomena of light scattering and diffraction on butterfly scales^{24,36}. The wings of butterflies and moths have various

structures, color patterns and types of scales which requires in the study focusing on relatively small parts of the wing, e.g. eyespot³⁷. There are scientific reports that the wing scales ultrastructure shapes the pattern of the reflectance spectrum even in closely related species³⁸. We suppose that the microscopic spectral characteristics of selected areas of the *D. chrysitis* and *D. stenochrysis* wing is worth exploring over a wide spectrum and may provide information useful in determining these species.

Remote sensing methods produce many features (spectral wavelengths), which means that rigorous numerical analysis is an indispensable step of the analytical procedure. Numerical analysis is necessary with all data sources, including genetics, RGB photos, and detailed hyperspectral measurements of various morphological characteristics such as shape, color, and wing pattern^{39–41}. Works also have been applied to determine sex, age, color diversity or taxonomical differentiation, but only a few of them report a truly successful classification^{42,43}.

A common feature of “sibling species” is that the visible differences between individuals assigned to one species may be more distinct than the differences between species themselves⁴⁴. Surprisingly, no attempt has yet been made to distinguish species complexes of Lepidoptera by studying parts of a wing using the full spectral range and advanced numerical techniques. In our work, we present the experiment results, comparing the detailed expert classification of *D. chrysitis* \ *D. stenochrysis* with results of the chemometric analysis of the full range of reflectance spectrum between 400 and 2100 nm, obtained under the microscope from two distinct groups of scales on the forewing which have completely different melanin content.

Methods

Specimen sampling and depository.

The research was carried out on 95 individuals of *D. chrysitis* (43) and *D. stenochrysis* (52). Male and female imagines were caught in light traps between 1995 and 2018 in Poland. Legislative and determinative features, including the location and trapping date of *D. chrysitis* and *D. stenochrysis* individuals, are presented in Supplementary Table S1. All sampled specimens were deposited in Department of Entomology and Environmental Protection of the Poznań University of Life Sciences.

In Poland, the populations of eastern and western moths occur sympatrically, constantly mixing, and therefore, we can exclude that the moths came from a few homogeneous populations. This is often a problem in studies involving relatively small areas for species that extend over much of the Palearctic. In addition, moths have been collected over a period of more than 20 years, which further diversifies the risk of “selective choice”. For these reasons we examined the potential influence of the age of specimens in collection on the quality of discrimination of species. Similarly, we assessed whether male and female individuals differ spectrally. For this purpose, we used the same discriminant analyzes described later in the paper. We notice that currently we cannot train successful learners predicting neither age-based fading nor sex of specimen. Also, the full effectiveness of the RF classifier (described in section Results)

in separating species indicates that those two properties does not affect the features favored by the classification models.

Expert classification.

The species of each individual was determined on the basis of following features: morphology of genitalia and colorization of the front pair of wings. In a laboratory, the body parts and the external genitalia were dissected in a standard way for each individual. The abdomen was first removed and dipped for 24–36 hours in 10% caustic potash (KOH). Genitalia were then removed from the softened surrounding tissues. The aedeagus was removed, and the external genitalia was partially dehydrated with ethanol and mounted on glycerine between the microscope slides and cover slips. Because of their three-dimensional shape and fragile structure, the endophalli were stored in liquid glycerine. The species determination in the collection described above was carried out by entomologist, professionally dealing mainly with moths from the Noctuidae family. Latest available publication of Ronkay et al.², and also the drawings of the genitalia shown in the several papers served as a key to species determination^{3,4,45}.

The coloring of the wings is an ambiguous feature. *D. chrysitis* differs externally from *D. stenochrysis* by its unbroken brown median area and more indistinct subterminal line². Depending on the individual, species traits may be very strong, strong or weak. The individuals with the typically marked species features on the forewings of *D. chrysitis* (Fig. 1a) and *D. stenochrysis* were classified as having very strong features. *D. chrysitis* (Fig. 1a) and *D. stenochrysis* (Fig. 1b) individuals with typically marked species features on the wings of the first pair were classified as having very strong features. The strong features of *D. chrysitis* butterflies typically had a species-specific not divided central field, but a sinuous subterminal line at which the metallic area ends, resembling the subterminal line characteristic of *D. stenochrysis*. Strongly featured *D. stenochrysis* individuals had a clearly divided brown median area by a band of golden iridescent scales over 1 mm wide. Individuals with weak features were marked on the basis of the structure of their genitals.

Expert analysis is always falsifiable. For this reason, the results of this part served as the ground truth reference for the spectral and chemometric procedures.

Spectral measurements.

Spectra were measured with a system consisting of the ASD FieldSpec 3 spectrophotometer (FieldSpec Analytical Spectral Devices, Inc., Boulder, Colorado, USA) attached by optical fiber to microscope NU 2 (VEB Carl Zeiss Jena, Jena, Germany). The spectrophotometer recorded the reflected electromagnetic radiation in the wavelength range from 350 to 2500 nm with a spectral sampling of 1.4 nm from 350 to 1000 nm and 2 nm from 1000 to 2500 nm. The spectral resolution in VIS was 3 nm and at 1400 and 2100 nm was 10 nm. The spectrophotometer was calibrated with a Spectralon (Labsphere) white standard before each measurement series. A plan apochromat objective (25x) and a coaxial illumination with a halogen lamp were used in the microscope.

Spectral measurements were carried out on the two dominant color areas on the wings, brown and golden iridescent on the dorsal side of forewing, which are typical of *D. chrysitis* and *D. stenochrysis*. In the case of the brown area the reflectance was measured from the median area at the front edge of the forewing while golden iridescent included subterminal area. In measurement location wings are covered with scales of two types. Cover scales are visible from the outside and brown ground scales below them (Supplementary Fig. S1). Brown area is covered with brown melanin-pigmented cover scales. While on the shimmering area, cover scales are actually colorless melanin-deprived and referred to as glass scales. As a result of light interference and diffraction, they form a physical color ranging from bluish to dominant gold to copper. Glass scales differ significantly in structure from brown scales, which are perforated and higher in cross-section (Supplementary Fig. S1 and S2). The base brown scales are also perforated. The scales are arranged in many layers on the wing.

The spectra at wavelengths below 400 nm and above 2100 nm exhibited high levels of noise, and they were not used in further analysis. The measurements were made in triplicate on the dorsal forewing of each of the 95 individuals and then the values have been averaged for each item. Results of spectra measurements are presented in Fig. 2.

Numerical analysis.

Spectral measurements provided 95 cases, each with what means 3402 potential predictors (two types of scales times 1701 bands). The collected dataset is small from a data science point of view, especially compared to the number of estimators. Such a situation is typical for most life science projects, where the costs limit the number of cases. In the first step, to highlight the spectra shape descriptors, data was transformed using Savitzky–Golay filter⁴⁶. The filter removes unnecessary effects like baseline shifts and noise resulting from the non-ideal sampling process. The SG filter requires three parameters and by experiments we found that configuration: differentiation order = 2, polynomial order = 2, and window size = 5 provides the best results. Data after SG transformation are input features for all classifiers. The entire procedure is summarized in Fig. 3.

Typical machine learning procedures require dividing the collected data into at least two sets: the training - used to train the model, and the testing set - used for an independent evaluation of the model accuracy. With a small number of cases and many features, the role of features in the trained model may vary significantly for various training and testing sets. The consequence of a small number of cases is the model's potential overfitting to the provided data⁴⁷. In our work, we attempt to reduce overfitting and detect essential predictors by two independent methods. The first is fully stochastic both at the sampling and training level and works with a complete spectrum. The second is fully deterministic, and it first selects a limited subset of potentially useful spectral features and then discriminates species by an exhaustive search inside this subset only. Both approaches have their own disadvantages, especially on small datasets. Stochastic approach, despite its utility may result in random, non-optimal configuration, difficult to reproduce. Deterministic approach, through dimensionality reduction processes may make it

difficult to find less obvious but more optimal solutions. For this reason, the species discrimination process will be controlled by two independent algorithms.

The first discriminating procedure uses Random Forest which is a stochastic algorithm and its results may change between each recurrence. Moreover, the standard procedure of machine learning training requires to hold part of the cases for the testing set. It means that the model is trained based on 60–80% of the entire dataset. Machine learning models gain their highest performance when the distributions of selected predictors (features) in the training and validation sets are close to each other. With large datasets this is not the case, but when the dataset is small, such an assumption is difficult to fulfill. It means that there is a risk that depending on the composition of the training set, the selected predictors will not be relevant for the cases included in the testing set. It is mainly related to high efficiency on the training set and low efficiency on the test set.

To determine the risk of overfitting, we applied a bootstrapping procedure, including 300 iterations. If such overfitting risk is high, we can expect that the testing set's performance will vary from low to high between iterations. If such a risk is low, each iteration shall return with similar performance. Moreover, if the trained model will select the same predictors at each iteration, we can expect that correct and incorrect classification will include the same cases each time. At each iteration, the entire dataset is randomly divided into the training set including 70% of cases and the testing set with remaining 30%. It means that we have 300 different training sets assessed by 300 different testing sets and each case had a chance to be included in the testing set 60 times on average. It is a large enough number to assess the stability of the process of discrimination. The RF algorithm also provides "feature importance" a Gini index describing how each variable decreases the impurity of RF internal splits. The index varies between 0 and 1, where 0 denotes that the variable cannot increase the purity, while 1 means that variable allows to split the dataset into pure subgroups.

In the second procedure we first attempted to define a group of the most significant spectrum bands for the distinction of species. As "the most important", we define a limited, possibly small number of predictors that successfully separate the species under investigation. First, all bands were sorted in descending importance order. We used the value of two-sample Kolmogorov-Smirnov (K-S) test as a variable importance index⁴⁸. The usefulness of this statistic to feature selection results from the absence of the assumption on the form of the distributions of compared sets. For each wavelength D-statistic was calculated using Eq. (1):

$$D_s = \sup |F_{dc}(x) - F_{ds}(x)| \quad (1)$$

where: F_{dc} and F_{ds} empirical distribution of *D. chrysitis* and *D. stenochrysis* subsets, and \sup is supremum function i.e. the choice of the largest operand value. D-statistic is 0 when two distributions fully overlap, values between 0 and 1 when partially overlap, and 1 if two distributions are completely disjoint. All wavelengths were sorted from the highest to lowest D-statistics, it means from the most to least separating features. Next, we searched for the minimal combination of predictors providing perfect separation between species. To evaluate the separation we used an accuracy of Linear Discriminant

Analysis (LDA), which is considered, as one of the best tools to find a minimal effective combination of spectral features⁴⁹.

Software.

We used the R programming language⁵⁰ for the analysis; the Savitzky-Golay filter from *prospectr*⁵¹ package to transform spectral curves; the *ranger*⁵² package to train the random forest models; linear discrimination from the *MASS*⁵³ package.

Results

Stochastic discriminant analysis.

In this procedure, all spectra, both of golden iridescent and brown fragments of the wing surfaces, were used simultaneously for classification. The classification achieved average compliance with the species classification made by entomologists at the level of $99\% \pm 1\%$. 91 individuals out of 95 were correctly classified during each of 300 testing iterations, and the remaining four specimens with over 98% correctness. It demonstrates the high stability of models trained against different training sets. Four cases where a match with expert classification was less than a hundred percent was observed. For *D. chrysitis* it was two cases, one recognized with weak and one with very strong confidence. For *D. stenochrysis* it was similar, one recognized with weak and one with strong confidence. This indicates that the expert self-confidence does not translate into the performance of RF classifiers because expert analysis and our method use a different set of characteristics and thus spectrometric properties are more stable than phenotypical features.

We also calculated the Kendall's rank correlation between expert confidence (Supplementary Table S1) and classification performance at the individual level. The results were not statistically significant, with $\tau = 0.07$ at $p\text{-value} = 0.49$. A one hundred percent reproducibility on the validation set for almost all cases indicates the presence of a group of relatively strong predictors among 3402 bands. The predictors are effective regardless of the configuration of the training sample. For each iteration of the simulation, the values of the importance index of all variables were recorded. After performing all iterations, the average value of the indexes was calculated. For the top 20 variables, listed in Table 1, the index value ranged from 0.49 for G1767 (glass scale, wavelength 1767 nm) to 0.22 for G1319 and only for the first 7 indicators the value of the index exceeded 0.3.

Deterministic discriminant analysis.

Figures 4 show the importance of each band by means of the D-statistic between species for given wavelength for glass and brown scales. D-statistics ranges from 0.07 to 0.83, but associate p-values show that only D-statistics larger than 0.4 allow to accept the null hypothesis that distributions inside species for the given band are different. This condition was true for no more than 500 bands. Only nine bands, all measured from the glass scales, have D-statistic greater than 0.7, with a maximum slightly

exceeding 0.82 for band 1378 nm. Twenty wavelengths with highest discriminating potential are also listed in Table 1. It is worth pointing out that 10 best wavelengths indicated by D-statistics are the same as 10 best resulting from the stochastic procedure, but in slightly different order. All these 10 wavelengths were measured from glass scales.

Table 1

List of 20 best wavelengths indicated by Random Forest feature importance and Kolmogorov-Smirnov D-statistics.

Method	Wavelengths: G - glass scales, B - brown scales
Random Forest	G1767, G1378, G1117, G1367, G1318, G1770, G1385, G714, G1110, G716, G1419, B1637, G1841, G1480, B1100, G1651, B1399, B1367, B1248, G1319
Kolmogorov-Smirnov D-statistics	G1378, G1767, G1367, G714, G1770, G1117, G1318, G716, G1110, G1385, B1637, B1367, G1841, G1419, G1480, G775, G1427, B716, G1533, B1535

There is no single spectral feature which fully distinguishes species. It means that we must search for a minimal number of bands from the entire spectrum whose linear combination allows such distinction. The discrimination process was undertaken by sorting features, starting from the most important. We found that linear combination of the first nine most important wavelengths (all from glass scales) perfectly differentiate species in LDA space, thus limiting the number of possible band combinations (Fig. 5). We also tested the potential of spectra derived from the brown scales only. The discriminatory potential of the individual wavelengths is smaller compared to those sensed from glass scales, but their linear combination already shows similar performance to the bands from brown scales (Fig. 5).

Minimal discriminatory combination.

In the last step of the analysis, we searched for the smallest possible group of wavelengths for each of the scales separately, which allow for full separation of both species. It should be emphasized that this is not an performance of LDA classifier measured on an independent testing set but only a group of bands for which there is a hyperplane in the multidimensional space completely separating *D. chrysitis* and *D. stenochrysis* in the collected dataset. There exist such four independent subsets for glass scales - one subset with three and three subsets with four wavelengths, and two subsets for brown scales - one with four and one subset with five wavelengths. All are presented in Table 2. Other, higher dimensional configurations are just a supersets of those mentioned above. In that way we can indicate the most important wavelengths - for glass scales there are: 1378, 1767, 716, 1385, 1117, 1318, nm and for brown: 1637, 1367, 894, 1942, 716, 1767 nm. Band 1535 nm is omitted from the brown scale list because it does not contribute to any subset that fully separates species. The isolation of the wavelengths of greatest diagnostic importance prompts us to insight into the biophysical and biochemical reasons for their relevance. The above thesis becomes all the more important because both the stochastic and deterministic analysis showed the existence of the same 10 most important bands for glass scales. Moreover, the LDA revealed the significance of two wavelengths (716, 1767 nm) in both brown and glass scales.

Table 2

The minimal combination of wavelengths to achieve full discrimination of species in LDA space. *Denote the D-statistic which is equivalent for a single band. The wavelengths marked in bold appear for the first time in the given subset.

Scales	Number of bands (dimensions)	List of bands (nm)	Accuracy
Glass	1	1378	0.826*
	2	1767, 716	0.968
	3	1385 , 1767, 716	1
	4	1378, 1767, 716, 1117 1378, 1767, 716, 1318 1378, 1318 , 716, 1385	1
Brown	1	1637	0.695*
	2	1637, 1535	0.926
	3	1637, 1367, 894 1637, 1367, 1942	0.968
	4	1637, 1367, 1942, 894	1
	5	1637, 1367, 1942, 716, 1767	1

Discussion

Previous studies on the spectral features of wings in moths of the genus *Diachrysia* focused on explaining the phenomenon of golden iridescence in *D. chrysis*^{54,55}. This arises as a result of interference, scattering and absorption in the structures of the glass scales. Based on our results we can apply the same for *D. stenochrysis*, whose wings have not been the subject of spectral studies so far. It should be noted that in the visible range there is a large variation in the iridescent color formed on colorless scales in the population of both species, from pale blue to pale copper. These color forms of moth wings were distinguished at the beginning of the 20th century⁵⁶. Using the reflected light spectrum recorded in the VIS-SWIR range, we have collected other, potentially more useful information than that provided by the visible range. The light has a relatively low ability to penetrate the inner wing structures under a layer of scales and the radiation is strongly reflected. Based on that, we recognized that the spectrum recorded by the spectrometer was shaped by the scales structures on the upper side of the wing.

Our approach to species identification based on spectral characteristics of spotty-defined areas of the wing and machine learning is original and new. So far, microscopic spectra studies have focused on the

identification of phenomena and comparative characteristics of wing scales^{24,36}. The analysis of the spectra is so complicated that even an trained expert can have problems distinguishing between these species without using chemometric methods. In our case, machine learning was crucial for decoding the information written in the form of the reflectance spectrum and to isolate information of significant importance for the species determination of *Diachrysia* siblings. Comprehensive comparison between expert classification and iteratively trained stochastic RF classifiers revealed that models very rarely incorrectly assign labels to validation data, regardless of the changes of the training sets composition. This indicates that the diagnostic features favored by RF repeats in all individuals of the collection. It was also confirmed by Kolmogorov-Smirnov D-statistics which indicated the same group of predictors. This means that with the optimal size and quality of the training set, spectral measurements coupled with machine learning can be an effective tool not only for mass species differentiation but also can provide insight into the reasons for this variation.

Since the identification of insects based on chemometric analysis is a new issue, an applicable spectral measurements library has not yet been developed. The only source of knowledge about the relationship between spectral bands and the properties of scales are previous studies. However, in-depth study of Lepidoptera wing scales that would explain the phenomena behind the meaning of the spectral features does not yet exist. In the developed model, the SWIR range turned out to be the most important in terms of species diagnostics. This is also seen in Fig. 6 where the visual differences between species spectra are clearly visible. In the optimal model we have three or four unique bands subsets for glass scales and four or five unique subsets for brown (Table 2). Few bands are of particular importance: 1767, 716 nm and located in close proximity 1378/1367 nm appear in unique sets, both in glass and brown scales. Those three bands suggest the presence of features important for the separation of moth species and independent of the type of scales. The remaining bands are important mainly for one type of scales. The amplitude of the spectra near these bands is small, indicating small differences between the two species.

The skeleton of the butterfly's scales is made of a composite material. There are three main components of Lepidoptera wing scales: chitin, proteins and melanin⁵⁷. The overall characteristic of the spectrum of *Diachrysia* scales, especially glass scales, is very similar to the general spectrum of chitin presented by Apetroaei et al.⁵⁸. Glass scales are colorless and melanin-deprived⁵⁹. Thus, chitin in glass scales seems to be the most important factor differentiating the siblings *Diachrysia*. And although the light reflected from the surface covered with glass scales may carry a weak signal coming from the brown ground scales below. We should also note that in the range of the spectrum we analyze, the protein components of the scales are not identified, although they may mask other signals. There are studies on the insects cuticular chitin spectral characteristics and its role in insect detection^{28,60,61}. Two significant wavelengths for glass scales: 1767 and 716 nm were repeated for brown scales, so we claim to be chitin dependent. Based on Schröder-Turk et al.⁶² we assume that one of the factors differentiating *Diachrysia* siblings may therefore be qualitative composition and the chiral structure of the chitin molecules. However, the main differentiating factor for both *Diachrysia* is ultrastructure of the scales, which significantly influences the shape of reflectance spectrum between butterfly species^{36,38}. Studies on structural color analysis proved

that the chitinous structure undoubtedly shapes the reflectance spectra which can be estimated from it^{63,64}. Azofeifa et al.⁶⁴ observed that with the increasing complexity of the structures of the analyzed chitin material, the simulated spectra showed weakening of the wave pattern. Thus, not only the overall shape of the curve but also the subtle changes in monotonicity of the curve depend on the complexity of chitin layers. Figure 6 presents location of the selected wavelengths indicated in “Minimal discriminatory combination” section of “Results” against medians of the raw spectra. All values correspond to the location of changes in monotonicity of the spectral curves. This means that wavelengths relate to subtle physicochemical differences between scales of *D. chrysitis* and *D. stenochrysis*, rather than the qualitatively identified presence or absence of selected chemical components.

Recently, scale ultrastructure can be described numerically on the basis of scanning electron microscope studies as shown by Day et al.³⁷. A detailed understanding of this structure could help answer the question that bothers us about the biophysical basis of the separateness of *Diachrysia* siblings. Scales ultrastructure is biochemically dependent on the remaining components of the scales. Melanin is the significant factor influencing the brown scales ultrastructure and response of our model. Moths' wings are usually dull as a consequence of high melanin content in brown or dark scales²⁵. The same we can state on brown melanin-pigmented scales on *Diachrysia* wings. There exists several types of melanin and it strongly influences the color and ultrastructure of moth scales⁶⁵. Brown scales on *Diachrysia* wings have a different spatial structure compared to glass ones (Supplementary Fig. S1 and S2). Perforated and pigmented brown scales scatter and absorb light more intensively than non-perforated⁶⁶. Thus melanin, being one of the biochemical determinants of the formation of scale ultrastructure, may therefore indirectly influence its differentiation between siblings moths. Although we cannot clearly explain the role of all the indicated wavelengths, we consider that the features denoted as important for species separation relate to the scale ultrastructure and its major components, chitin and melanin. Moreover, it should be emphasized that two wavelengths: 1767 and 716 nm, were identified in both types of scales, which minimizes the risk that the indicated wavelengths are random.

Conclusions

Biologists dealing with systematics and evolution of Lepidoptera are constantly looking for new sources of characteristics and methods to distinguish taxa. Our approach is the opposite of expert analysis, where the first step is to identify important features and then make a decision as to species affiliation. The novelty of our approach is that the combination of reflectance spectroscopy and machine learning has been used to separate hardly distinguishable species: *D. chrysitis* and *D. stenochrysis*. We used microscopic spectroscopy to obtain pure spectra of two distinct groups of scales on the forewing. Applied methods of species discrimination proved a decent accuracy and time savings, especially compared to the expert approach. The advantage of the RF is a simplicity of procedure and no preliminary assumptions about the importance and distribution of individual predictors. Especially the latter means that none of the essential features will be omitted in the decision-making process. The

feature selection procedure allows indicating a relatively small set (between three to six) of predictors with the highest discriminatory potential.

Presented paper proved that spectral investigation of scales can be good and low-cost discrimination of selected moth species. Spectral differences between *Diachrysia* siblings are mainly the result of ultrastructure of scales influenced by chitin and melanin, the nature of which is not achievable for spectral studies. The differences are subtle and quantitative, nonetheless statistically significant, and allow, using just several predictors, to fully discriminate these two species. The role of chitin in the discrimination of siblings *Diachrysia* is highlighted by the wavelengths separated for melanin-deprived glass scales. Two of these wavelengths of 1767 and 716 nm appear in unique sets, both in glass and brown scales. Since chitin-melanin cuticular skeleton occurs in Lepidoptera scales and are generally common in the world of insects, the proposed method may be useful and effective for the determination of other species of Lepidoptera.

Declarations

Data availability

The input data (legislative data and spectral measurements) with a description, and the results of this study are available in the GitHub repository (<https://github.com/kadyb/diachrysia-classification>) under the MIT license.

Code availability

The fully functional and reproducible codes have been published publicly in the GitHub repository (<https://github.com/kadyb/diachrysia-classification>) under the MIT license.

Acknowledgements

We would like to thank Mr. Jerzy Wrocławski, an expert in optics, for integrating the spectroscope with the microscope for the purposes of our research.

Author contributions

H.R., R.W., J.P. developed the research concept. K.D. was responsible for the mathematical part, solved the research problem, wrote the code and prepared figures. R.W. collected individuals and made entomological evaluation. J.P., H.R. made spectral measurements under a microscope. J.J. described the methodological part, supervised the chemometric analysis and plotting the graphical results. K.D., H.R., J.J., J.P., R.W. wrote the main manuscript text. EG, MG took pictures with an electron microscope imaging technique. All authors contributed to the critical appraisal of the paper and approved the final version.

Additional information

Correspondence and requests for materials should be addressed to H. R. (henryk.ratajkiewicz@up.poznan.pl) or J. J. (jaroslaw.jasiewicz@amu.edu.pl).

Competing interests. The authors declare no competing interests.

References

1. Fibiger, M. *Noctuinae I*. (Entomological Press, 1990).
2. Ronkay, L., Ronkay, G. & Behounek, G. *Plusiinae 1*. (Heterocera Press, 2008).
3. Goater, B., Ronkay, L., Fibiger, M. & Goater, B. *Catocalinae & Plusiinae*. vol. 10 (Entomological Press, 2003).
4. Kostrowicki, S. Studies on the palaearctic species of the subfamily Plusiinae (Lepidoptera, Phalaenidae). *Acta Zool Cracov* **10**, 367–472 (1961).
5. Lempke, B. Notes on the genus *Autographa* HÜBNER, subgenus *Chrysaspidia* HÜBNER (Lep., Noctuidae). *Entomol. Ber.* **26**, 64–71 (1966).
6. Urbahn, E. Zur Artenfrage *Plusia chrysitis* L. - tutti Kostrowicki (Lepidoptera, Noctuidae). *Reichenbachia* 129–136 (1966).
7. Urbahn, E. Zur Klärung der *Plusia chrysitis*-tutti-Frage durch Eizuchtuntersuchungen (Lepidoptera, Noctuidae). *Reichenbachia* 133–137 (1967).
8. Rezbanyai-Reser, L. *Diachrysia chrysitis* (Linnaeus, 1758) und tutti (Kostrowicki, 1961) in der Schweiz. Ergebnisse von Pheromonfallenfangen 1983-84 sowie Untersuchungen zur Morphologie, Phanologie, Verbreitung und Oekologie der beiden Taxa (Lepidoptera, Noctuidae: Plusiinae). *Mitteilungen Schweiz. Entomol. Ges. Bull. Soc. Entomol. Suisse* **58**, 345–372 (1985).
9. Priesner, E. Artspezifische suxuallockstoffe für Männchen von *Diachrysia chrysitis* (L.) und *D. tutti* (Kostr.) (Lepidoptera, Noctuidae: Plusiinae). *Mitt Schweiz Entomol Ges* (1985).
10. Bruun, H. H. Longitudinal ridge density of hind-wing scales of *Diachrysia chrysitis* (L.) and *D. tutti* (Kostr.) captured with pheromones (Lepidoptera, Noctuidae). *Entomol* **67**, 125–127 (1987).
11. Svensson, I., Douwes, P. & Stille, B. Are *Diachrysia chrysitis* (L.) and *D. tutti* (Kostrowicki) different species? (Lepidoptera: Noctuidae). *Insect Syst. Evol.* **20**, 15–22 (1989).
12. Löfstedt, C. *et al.* Pheromone differences between sibling taxa *Diachrysia chrysitis* (Linnaeus, 1758) and *D. tutti* (Kostrowicki, 1961) (Lepidoptera: Noctuidae). *J. Chem. Ecol.* **20**, 91–109 (1994).
13. Inomata, S.-I., Watanabe, A., Nomura, M. & Ando, T. Mating Communication Systems of Four Plusiinae Species Distributed in Japan: Identification of the Sex Pheromones and Field Evaluation. *J. Chem. Ecol.* **31**, 1429–1442 (2005).
14. Hille, A., Miller, M. A. & Erlacher, S. DNA sequence variation at the mitochondrial cytochrome oxidase I subunit among phenotypes of the sibling taxa *Diachrysia chrysitis* and *D. tutti* (Lepidoptera: Noctuidae). *Zool. Scr.* **34**, 49–56 (2005).

15. Huemer, P. & Hebert, P. DNA-Barcoding der Schmetterlinge (Lepidoptera) Vorarlbergs (Österreich) - Erkenntnisse und Rückschlüsse. *Inatura - Forsch. Online* **15**, 1–36 (2015).
16. Hausmann, A. *et al.* Now DNA-barcoded: the butterflies and larger moths of Germany. *Spixiana* **34**, 47–58 (2011).
17. Hopkins, G. W. & Freckleton, R. P. Declines in the numbers of amateur and professional taxonomists: implications for conservation. *Anim. Conserv.* **5**, 245–249 (2002).
18. Asiah Hassan, S. N., Abdul Rahman, N. N. S., Zaw Htike, Z. & Lei Win, S. Vision Based Entomology: A Survey. *Int. J. Comput. Sci. Eng. Surv.* **5**, 19–32 (2014).
19. Johnson, J. B. & Naiker, M. Seeing red: A review of the use of near-infrared spectroscopy (NIRS) in entomology. *Appl. Spectrosc. Rev.* **55**, 810–839 (2020).
20. Hennessy, A., Clarke, K. & Lewis, M. Hyperspectral Classification of Plants: A Review of Waveband Selection Generalisability. *Remote Sens.* **12**, 113 (2020).
21. Schultz, T. D. & Hadley, N. F. Structural Colors of Tiger Beetles and Their Role in Heat Transfer through the Integument. *Physiol. Zool.* **60**, 737–745 (1987).
22. Tsai, C.-C. *et al.* Physical and behavioral adaptations to prevent overheating of the living wings of butterflies. *Nat. Commun.* **11**, 551 (2020).
23. Yoshioka, S., Nakano, T., Nozue, Y. & Kinoshita, S. Coloration using higher order optical interference in the wing pattern of the Madagascan sunset moth. *J. R. Soc. Interface* **5**, 457–464 (2008).
24. Wilts, B. D., Vey, A. J. M., Briscoe, A. D. & Stavenga, D. G. Longwing (*Heliconius*) butterflies combine a restricted set of pigmentary and structural coloration mechanisms. *BMC Evol. Biol.* **17**, 226 (2017).
25. Stavenga, D. G., Wallace, J. R. A. & Warrant, E. J. Bogong Moths Are Well Camouflaged by Effectively Decolourized Wing Scales. *Front. Physiol.* **11**, 95 (2020).
26. Kaya, Y. & Kayci, L. Application of artificial neural network for automatic detection of butterfly species using color and texture features. *Vis. Comput.* **30**, 71–79 (2014).
27. Krishna, A. *et al.* Infrared optical and thermal properties of microstructures in butterfly wings. *Proc. Natl. Acad. Sci.* **117**, 1566–1572 (2020).
28. Ridgway, C. & Chambers, J. Detection of External and Internal Insect Infestation in Wheat by Near-Infrared Reflectance Spectroscopy. *J. Sci. Food Agric.* **71**, 251–264 (1996).
29. Dowell, F. E., Throne, J. E., Wang, D. & Baker, J. E. Identifying Stored-Grain Insects Using Near-Infrared Spectroscopy. *J. Econ. Entomol.* **92**, 165–169 (1999).
30. Stuart-Fox, D., Newton, E. & Clusella-Trullas, S. Thermal consequences of colour and near-infrared reflectance. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20160345 (2017).
31. Wilson, R. H., Nadeau, K. P., Jaworski, F. B., Tromberg, B. J. & Durkin, A. J. Review of short-wave infrared spectroscopy and imaging methods for biological tissue characterization. *J. Biomed. Opt.* **20**, 030901 (2015).
32. Song, H.-S., Lee, K.-T., Park, S.-M., Kang, O.-J. & Cheong, H.-S. Measurement of Deproteinization and Deacetylation of Chitin and Chitosan by Near Infrared Spectroscopy. *Korean J. Fish. Aquat. Sci.* **36**,

88–93 (2003).

33. Gebru, A., Brydegaard, M., Rohwer, E. & Neethling, P. Probing insect backscatter cross section and melanization using kHz optical remote detection system. *J. Appl. Remote Sens.* **11**, 016015 (2017).
34. Gebru, A. *et al.* Multiband modulation spectroscopy for the determination of sex and species of mosquitoes in flight. *J. Biophotonics* **11**, e201800014 (2018).
35. Brydegaard, M. & Svanberg, S. Photonic Monitoring of Atmospheric and Aquatic Fauna. *Laser Photonics Rev.* **12**, 1800135 (2018).
36. Trzeciak, T. M., Wilts, B. D., Stavenga, D. G. & Vukusic, P. Variable multilayer reflection together with long-pass filtering pigment determines the wing coloration of papilionid butterflies of the nireus group. *Opt. Express* **20**, 8877 (2012).
37. Day, C. R., Hanly, J. J., Ren, A. & Martin, A. Sub-micrometer insights into the cytoskeletal dynamics and ultrastructural diversity of butterfly wing scales. *Dev. Dyn.* **248**, 657–670 (2019).
38. Parnell, A. J. *et al.* Wing scale ultrastructure underlying convergent and divergent iridescent colours in mimetic *Heliconius* butterflies. *J. R. Soc. Interface* **15**, 20170948 (2018).
39. Zhu, L.-Q. & Zhang, Z. Auto-classification of insect images based on color histogram and GLCM. in *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery* 2589–2593 (IEEE, 2010). doi:10.1109/FSKD.2010.5569848.
40. Lim, J., Cho, J., Nam, T. & Kim, S. Development of a classification algorithm for butterflies and ladybugs. in *TENCON 2006 - 2006 IEEE Region 10 Conference* 1–3 (IEEE, 2006). doi:10.1109/TENCON.2006.344144.
41. Yang, H. *et al.* Research on insect identification based on pattern recognition technology. in *2010 Sixth International Conference on Natural Computation* 545–548 (IEEE, 2010). doi:10.1109/ICNC.2010.5583156.
42. Lazzari, S. M. N., Ceruti, F. C., Rodriguez-Fernandez, J. I., Opit, G. & Lazzari, F. A. Intra and interspecific variation assessment in Psocoptera using near spectroscopy. *Julius-Kühn-Arch.* **425** 139-144 (2010) doi:10.5073/JKA.2010.425.250.
43. Wu, S. *et al.* Artificial intelligence reveals environmental constraints on colour diversity in insects. *Nat. Commun.* **10**, 4554 (2019).
44. Ridley, M. *Evolution*. (Blackwell Pub, 2004).
45. Nowacki, J. *The Noctuids (Lepidoptera, Noctuidae) of Central Europe*. (Slamka, 1998).
46. Savitzky, A. & Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **36**, 1627–1639 (1964).
47. Tsamardinos, I., Greasidou, E. & Borboudakis, G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Mach. Learn.* **107**, 1895–1922 (2018).
48. Ivanov, A. & Riccardi, G. Kolmogorov-Smirnov test for feature selection in emotion recognition from speech. in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5125–5128 (IEEE, 2012). doi:10.1109/ICASSP.2012.6289074.

49. Piekarczyk, J., Ratajkiewicz, H., Jasiewicz, J., Sosnowska, D. & Wójtowicz, A. An application of reflectance spectroscopy to differentiate of entomopathogenic fungi species. *J. Photochem. Photobiol. B* **190**, 32–41 (2019).
50. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2021).
51. Stevens, A. & Ramirez-Lopez, L. *An introduction to the prospectr package*. (2020).
52. Wright, M. N. & Ziegler, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J. Stat. Softw.* **77**, (2017).
53. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, 2002).
54. Pantelić, D. *et al.* Scattering-enhanced absorption and interference produce a golden wing color of the burnished brass moth, *Diachrysia chrysitis*. *Phys. Rev. E* **95**, 032405 (2017).
55. Savić-Šević, S., Pantelić, D., Jelenković, B., Salatić, B. & Stojanović, D. V. Golden moth-inspired structures with a synergistic effect of interference, absorption and scattering. *Soft Matter* **14**, 5595–5603 (2018).
56. Romaniszyn, J. & Schille, F. *Fauna motyli Polski*. vol. 6 (1929).
57. Hunt, S. Composition of scales from the moth *Xylophasia monoglypha*. *Experientia* **27**, 1030–1031 (1971).
58. Apetroaei, M. *et al.* Improved method of chitosan extraction from different crustacean species of Romanian Black Sea coast. *UPB Sci. Bull. Ser. B Chem. Mater. Sci.* **79**, 25–36 (2017).
59. Gomez, D. *et al.* *Transparency in butterflies and moths: structural diversity, optical properties and ecological relevance*. <http://biorxiv.org/lookup/doi/10.1101/2020.05.14.093450> (2020)
doi:10.1101/2020.05.14.093450.
60. Dowell, F. E., Throne, J. E. & Baker, J. E. Automated Nondestructive Detection of Internal Insect Infestation of Wheat Kernels by Using Near-Infrared Reflectance Spectroscopy. *J. Econ. Entomol.* **91**, 899–904 (1998).
61. Kaya, M., Bitim, B., Mujtaba, M. & Koyuncu, T. Surface morphology of chitin highly related with the isolated body part of butterfly (*Argynnis pandora*). *Int. J. Biol. Macromol.* **81**, 443–449 (2015).
62. Schröder-Turk, G. E. *et al.* The chiral structure of porous chitin within the wing-scales of *Callophrys rubi*. *J. Struct. Biol.* **174**, 290–295 (2011).
63. Vigneron, J. P., Colomer, J.-F., Vigneron, N. & Lousse, V. Natural layer-by-layer photonic structure in the squamae of *Hoplia coerulea* (Coleoptera). *Phys. Rev. E* **72**, 061904 (2005).
64. Azofeifa, D. E., Arguedas, H. J. & Vargas, W. E. Optical properties of chitin and chitosan biopolymers with application to structural color analysis. *Opt. Mater.* **35**, 175–183 (2012).
65. Matsuoka, Y. & Monteiro, A. Melanin Pathway Genes Regulate Color and Morphology of Butterfly Wing Scales. *Cell Rep.* **24**, 56–65 (2018).
66. Stavenga, D. G., Stowe, S., Siebke, K., Zeil, J. & Arikawa, K. Butterfly wing colours: scale beads make white pierid wings brighter. *Proc. R. Soc. Lond. B Biol. Sci.* **271**, 1577–1584 (2004).

Figures



Figure 1

The dorsal part of a wing of *Diachrysia chrysitis* and *D. stenochrysis* and their forms with well-marked morphological features. a *D. chrysitis*. b *D. stenochrysis*

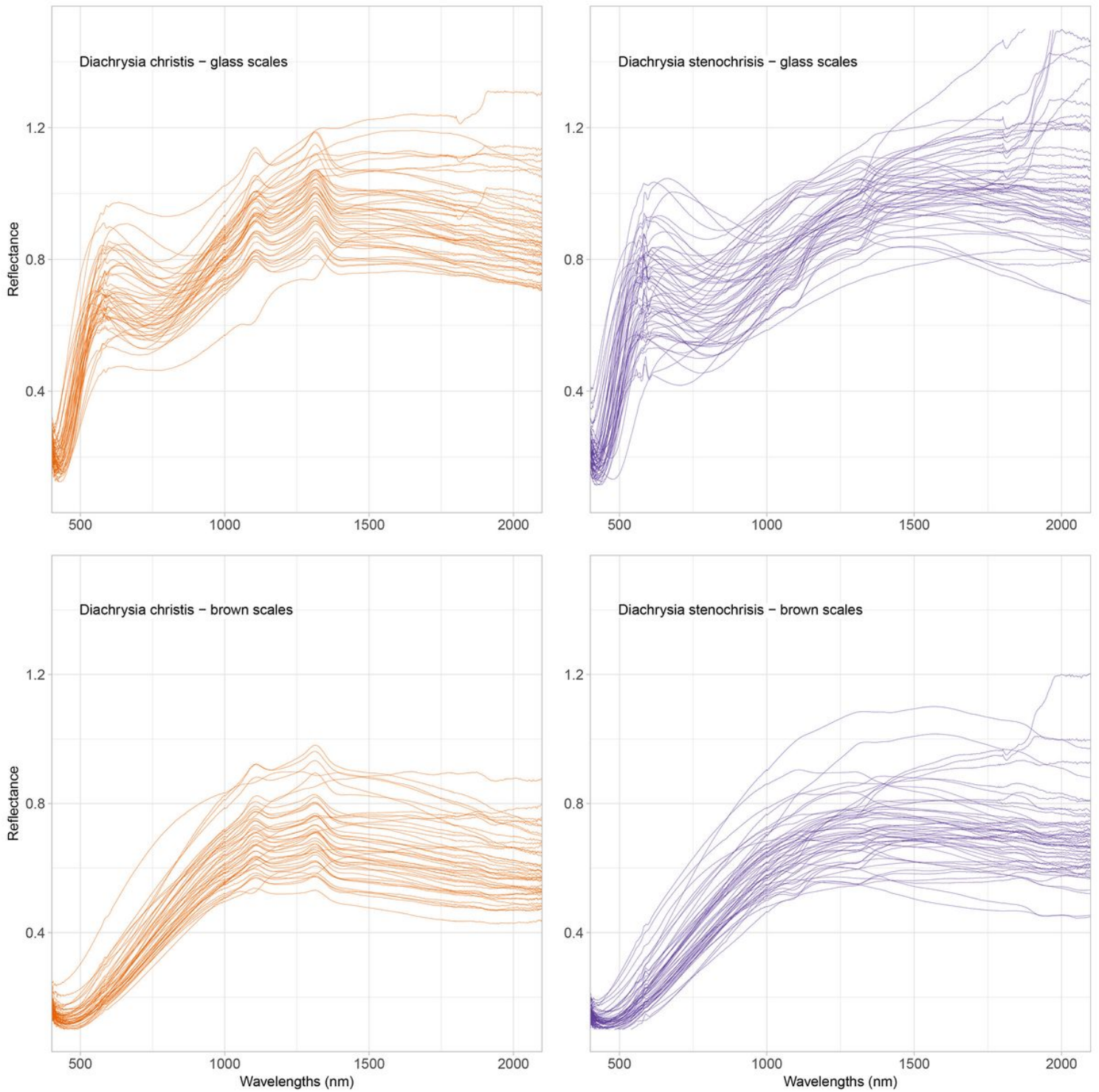


Figure 2

Raw spectra for *D. chrysis* and *D. stenochrisis* species.

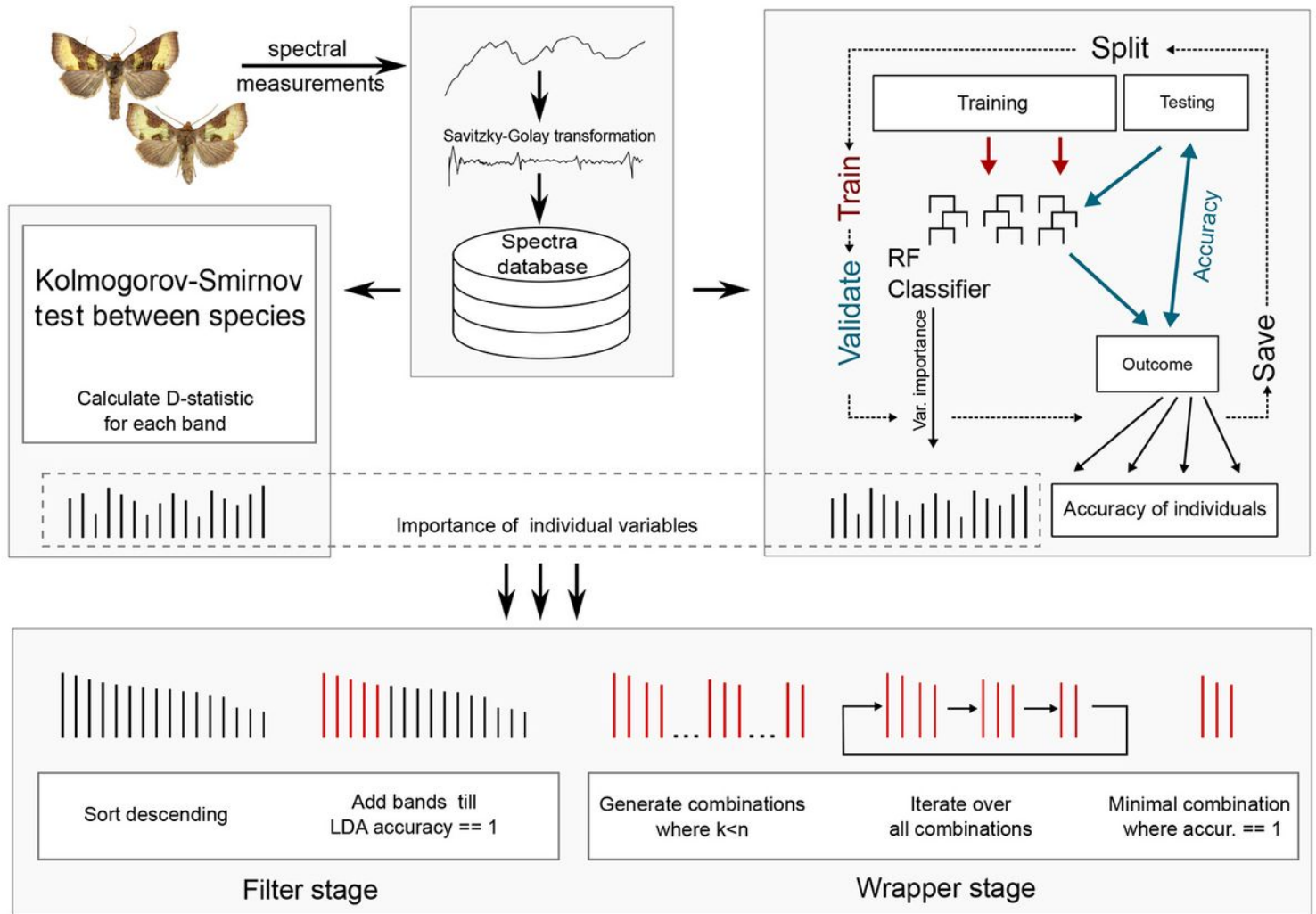


Figure 3

Block diagram of the procedure. A total of 300 iterations of model training and testing were performed, and the final results were averaged.

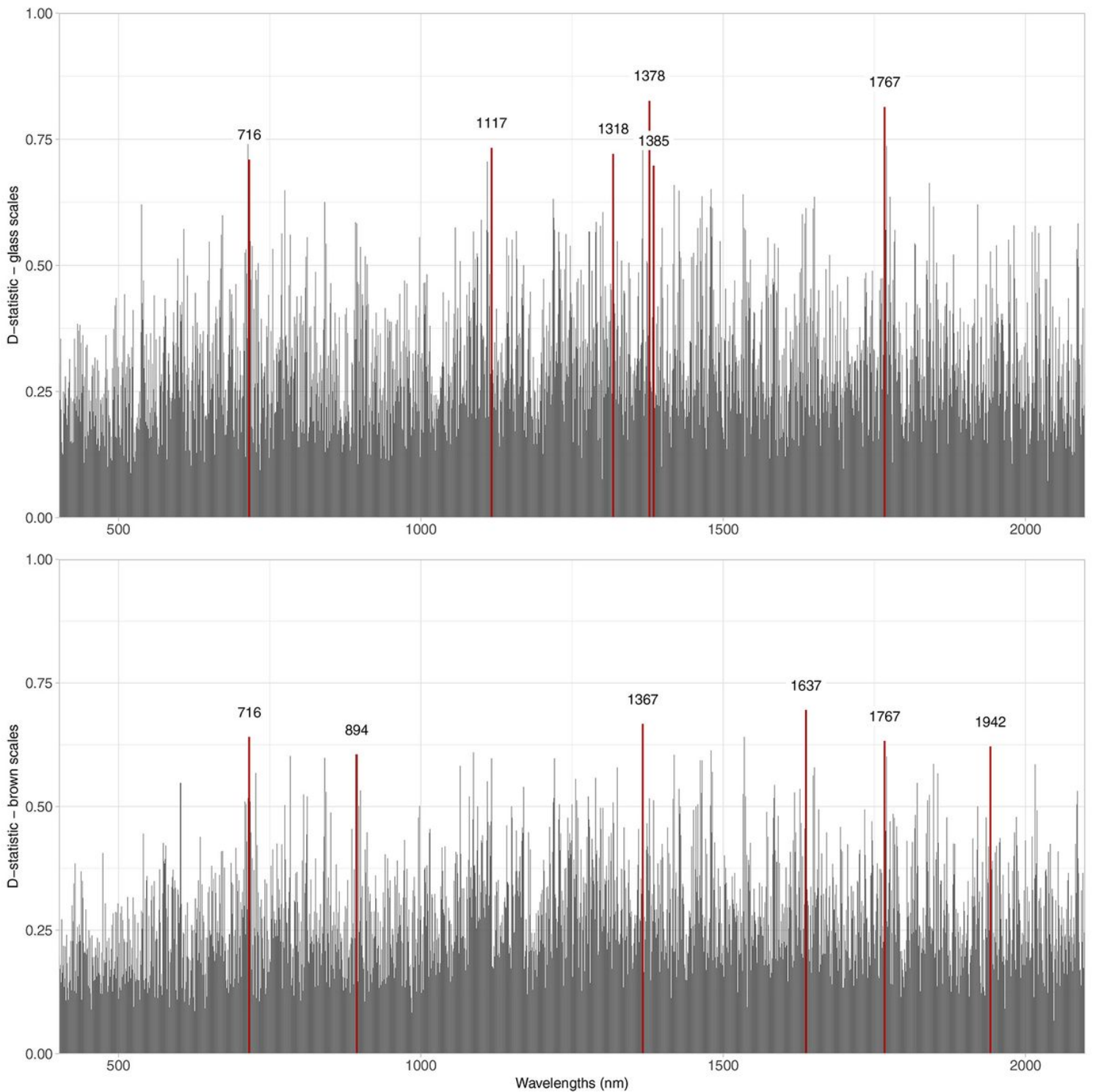


Figure 4

The importance of spectral bands for species classification. The higher the value, the greater the suitability for distinguishing between species. Most important bands are marked red. Notice that two bands 716 and 1767 nm appear in both types of scales.

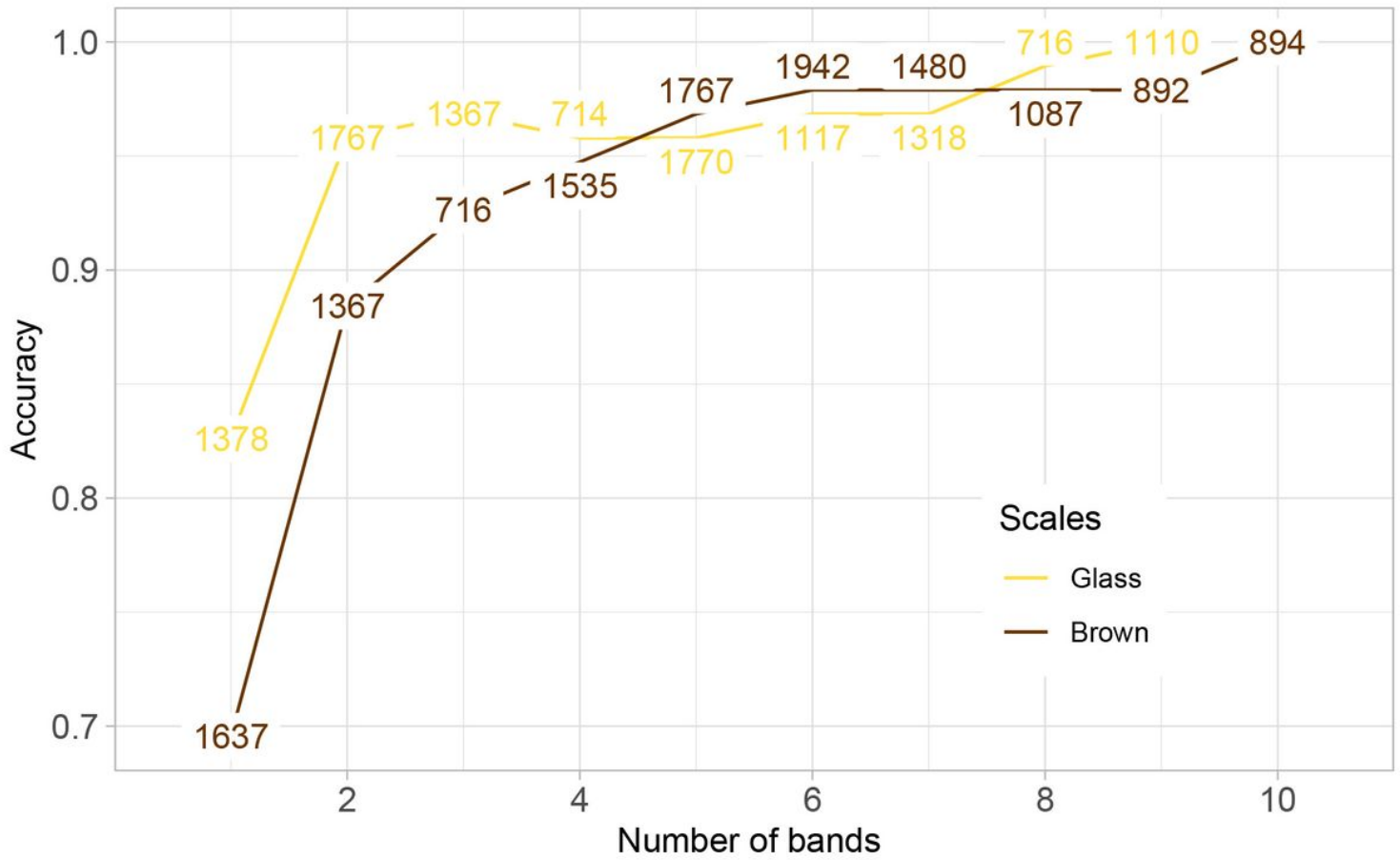


Figure 5

Relation of the LDA accuracy of species discrimination to the number of the consecutive spectral bands. Notice that the values on the Y axis do not start from 0.

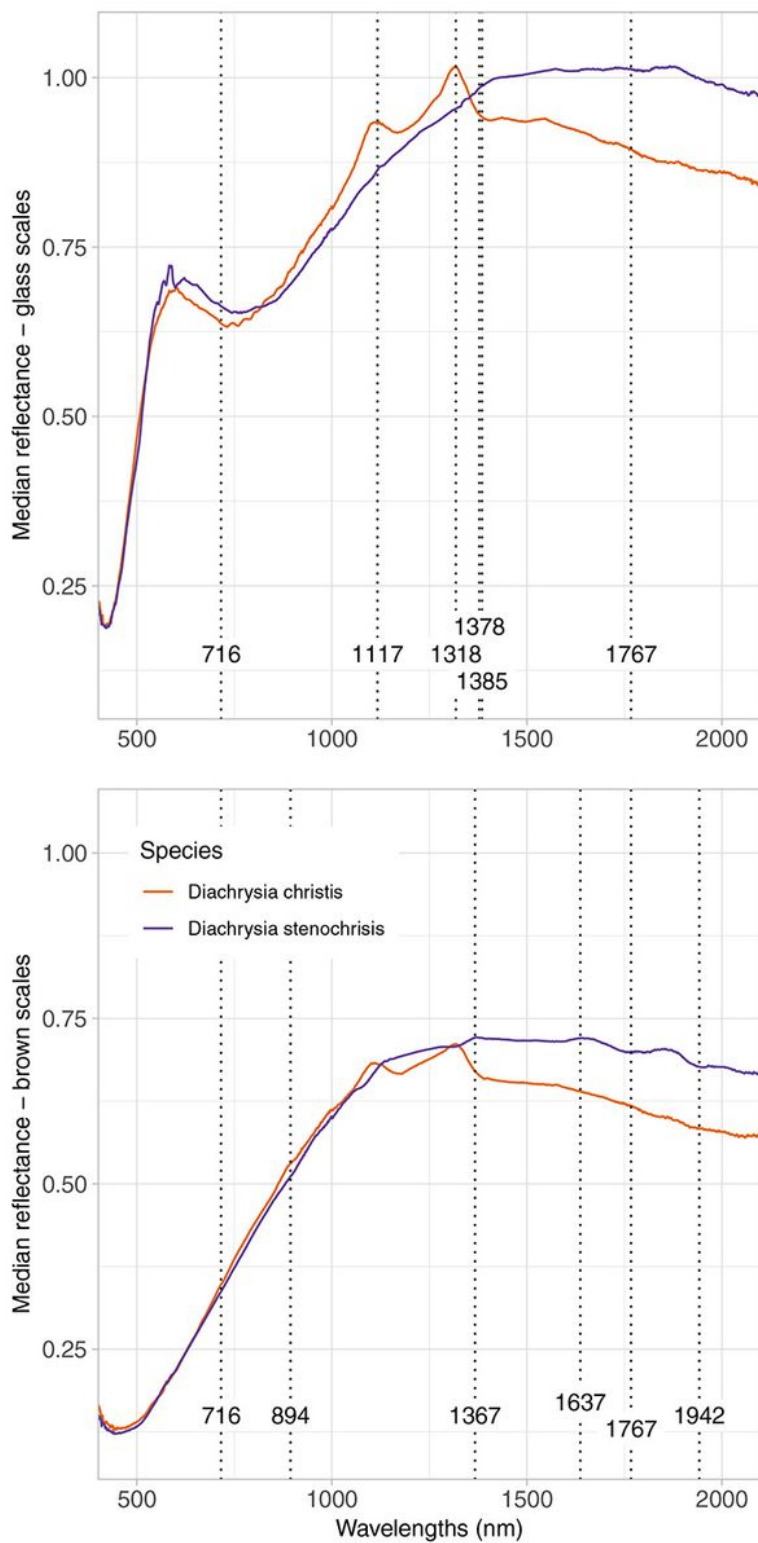


Figure 6

The medians of the species spectra for glass and brown scales with overprinted most important wavelengths.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryTableS1.xlsx](#)
- [SupplementaryFigureS1.jpg](#)
- [SupplementaryFigureS2.jpg](#)