scientific reports

OPEN

Physically-based modelling for retrospective detection of archaeological proxies (cropmarks)

Elias Gravanis✉ & Athos Agapiou

Identifying spectral anomalies such as cropmarks is a valuable approach to detecting concealed archaeological sites. In this study, we explore a physically based modeling strategy, interpreting cropmarks as stress-induced changes in vegetation, governed by radiative transfer dynamics. Measurements were obtained from a controlled test field during two campaigns conducted 13 years apart. Using PROSAIL in both forward and inverse modes, combined with statistical modeling, we construct synthetic spectral datasets to augment limited observations, thus facilitating robust training of classification models. An ensemble of machine learning algorithms is applied, with classifiers trained on synthetic data and refined via majority voting. The models achieve detection rates exceeding 90% on earlier observations, demonstrating effective retrospective application. Results highlight the influence of plant growth phase, with peak greenness data yielding stronger performance. Interestingly, injected noise improves robustness, albeit modestly. This study establishes a reproducible pipeline for archaeological prospection, merging physical simulation with machine learning. The integration of synthetic data offers a promising solution to data scarcity, and opens pathways for applying predictive models to archive aerial and satellite imagery, thereby potentially transforming remote sensing strategies in heritage research.

Keywords Archaeological prospection, Archaeological proxies, Cropmarks, Machine learning, Predictive modelling, Predictive accuracy

Detection of archaeological sub-surface remains through remote sensing sensors has been widely adopted in the past in different landscapes around the globe[1,2]. The detection of stressed vegetation due to the presence of shallow-buried remains, known in the literature as cropmarks, is considered as one of the most widely adopted methods[3–5]. These archaeological proxies are detectable mainly in arable land where barley, cereal or other crops are cultivated over shallow-buried remains. Cultivation may be either stressed (negative cropmarks) or enhanced (positive cropmarks), depending on the underlying context[6]. Such observations can be used as a first indication (proxy) of the existence of buried remains, helping researchers to prioritize their area of interest for field surveys.

Several examples and important discoveries have been made using this approach[7–9]. Satellite imagery, which records both visible and near-infrared wavelengths, can reveal spectral anomalies in crops through indices, enhancements, and other image-processing techniques. However, significant limitations highlighted by a series of studies indicate the complexity for capturing the cropmark phenomenon using remote sensing methods. The scientific community has worked in different directions by using integrated remote sensors[10–13], working with fusion models[14,15] or performing extensive studies with soil sampling and field campaigns[16–18]. Nonetheless, studies suggest that further research is needed so as remote sensing prospection methods to be more effective[16]. Indeed, until today, findings from targeted archaeological prospection campaigns are still regarded as isolated studies of local interest, limiting the advancement of this scientific field, i.e., developing regional models.

A path beyond case-specific studies lies in developing a deeper understanding of how cropmarks form and which factors govern their expression. The necessary theoretical advances are still missing in the literature, despite new data becoming available through technological developments in the space sector (i.e., higher spatial and spectral resolution sensors) and automation in image processing (i.e., deep learning algorithms). Such theoretical progress may arise naturally through the physical representation of cropmarks using theories of light propagation through vegetation. Radiative transfer models, such as PROSAIL[19], associate top of canopy reflectance spectral signatures with biophysical and structural canopy characteristics and are widely used for agriculture and forestry applications.

Department of Civil Engineering and Geomatics, Cyprus University of Technology, Limassol 3036, Cyprus. ✉email: elias.gravanis@cut.ac.cy

Previous work by the authors has followed this line of thought[20–23] constructing the first theoretical models for cropmark detection by combining ground-truth spectral data, radiative-transfer modelling, and statistical modelling. These approaches may eventually support regional-scale applications. These methods were further developed in[24] by utilizing machine learning techniques, to build models and assess their effectiveness in detecting cropmark signatures within a new measured set of signatures. In the present study we develop and assess retrospective predictive modelling, analyzing the ability of models built on present data to identify the presence of cropmark signatures in archived data. This context is particularly significant for discovering new archaeological sites, utilizing currently available datasets and subsequently extending these findings to archived information, and, in this way, exploring historic landscapes of interest that might not be visible nowadays.

Our ground truth data comes from reflectance signatures measured in the controlled environment of the Alampra test field during two campaigns, in the years 2011–2012[25] and 2025. This is a small test crop (barley) field, measuring 5 m×5 m, specifically constructed to simulate buried archaeological features and to provide clean top-of-canopy and soil data suitable for proof-of-concept analyses and modelling. Ground spectroradiometric measurements were taken during the mentioned seasons, where spectral signatures were collected over the "archaeological" and non-archaeological areas. A calibrated handheld GER 1500 spectroradiometer was used for this purpose for the 2011–2012 measurements and a FieldSpec HandHeld spectroradiomter was used for the 2025 measurements. Both handheld instruments can record the electromagnetic radiation between the visible to the near-infrared spectrum. The field spectral measurements were calibrated using a Lambertian spectralon panel.

The PROSAIL model provides a physical representation of measured signatures. Through PROSAIL inversion, each signature is associated with a set of leaf and canopy physical parameters, labeled as originating from an 'archaeological' spot (ARCHAEO or A) or a 'healthy' location (HEALTHY or H). The estimated parameter values are statistically modeled, and an empirical probability distribution is sampled to generate synthetic parameter values that closely reflect the estimated ones. Running PROSAIL in forward mode produces a set of simulated reflectance signatures, each carrying an A or H label. Machine learning techniques are then applied to model these synthetic datasets for predictive modeling. To leverage diverse algorithmic responses, we employ a range of classifiers. For robustness, we train the classifiers on batches of varying sample sizes, using majority voting to enhance reliability. This methodology was employed in[24] using the synthetic dataset based on the 2012 measurements[21,22] to assess the performance of models at detecting A signatures in the 2025 measurements. The results from that study showed a detection rate above 90%.

In the present study, we use such methods to construct the synthetic dataset based on the 2025 measurements to attempt to detect the A signatures within the 2012 measurements. Certain refinements of the previous methods were required. PROSAIL inversion is done by constraint-based regularization, minimizing a cost function that penalizes deviations from the midpoint of a chosen interval for each parameter (in accordance with the usual range for crops known from the literature). The 2025 measured dataset is sufficiently large to allow us to extend our analysis in this study, by modelling the spectral signatures for each observation date separately; that was not the case with the 2012 dataset, where we had to bundle all signatures together, and hence the simulated ones did not carry date label. Subsequently, the machine learning models based on the new measurements, also carry a date label. Hence, simulations and models correspond to a different growth phase of the crops and exhibit different predictive performance.

Materials and methods

Case study and in-situ campaigns

An experimental test area was established back in 2011 near the village of Alampra in Cyprus [UTM coordinates (WGS84, Zone 36 N): 535051 E, 3870818 N]. The site comprises a square plot of 5 m×5 m, designed to mimic ancient tomb structures buried at a depth of 25 centimeters. Throughout the preparation of this controlled experimental environment, careful measures were implemented to preserve the original soil stratigraphy, maintaining the natural soil surface at the top. The tomb structures were architecturally designed to closely replicate genuine archaeological buried features[25].

Following the site's establishment, the entire experimental area was cultivated with barley. During the crop growing season of 2011–2012, the researchers carried out more than sixteen spectroradiometric surveys over both the archaeological and non-archaeological sections of the study area. These campaigns involved the use of calibrated spectroradiometers: a GER 1500 device for the 2011–2012 season and a FieldSpec HandHeld spectroradiometer in 2025. Both instruments are sensitive within the visible to near-infrared spectrum (400–1,050 nm) and record across 512 distinct channels. These measurements were then resampled at 1 nm intervals using the instrument's add-on software. Although the analysis was conducted without applying FieldSpec→GER harmonization, based on the relative response filters (RSR) of the individual bands, a post-hoc evaluation using the GER-1500 band central wavelengths, its 3.2 nm full width at half maximum (FWHM), and a Gaussian spectral response function to convolve the FieldSpec spectra showed that the resulting absolute relative differences were consistently below 1%. Because these differences fall well within the GER-1500 radiometric calibration uncertainty (i.e. 5%), cross-sensor spectral mismatch is unlikely to influence the modelling outcomes.

To ensure accurate measurements of incoming solar radiation and reliable data collection, a calibrated Lambertian spectralon panel was employed. All spectral measurements were taken between 10:00–14:00 local time to minimize the effects of varying solar illumination. The field of view of the spectroradiometer was set at 4 degrees, covering roughly 0.02 square meters from a consistent height of 1.2 m. Initial radiance measurements were based on reference readings from the spectralon panel, followed by targeted data acquisition over selected locations of interest[19,20].

The analysis of[20] showed that the season January-February-March is the most relevant for modelling purposes. Hence a new campaign was conducted in the Alampra test field during this season in 2025. The

Date	Sample type		
	ARCHAEO	HEALTHY	SOIL
2025/01/20	50	50	51
2025/01/30	50	50	50
2025/02/08	50	50	50
2025/03/08	50	50	50
2025/03/12	50	50	50
2025/03/15	100	100	100

Table 1. Number of observed signatures per date and sample type in the 2025 campaign.

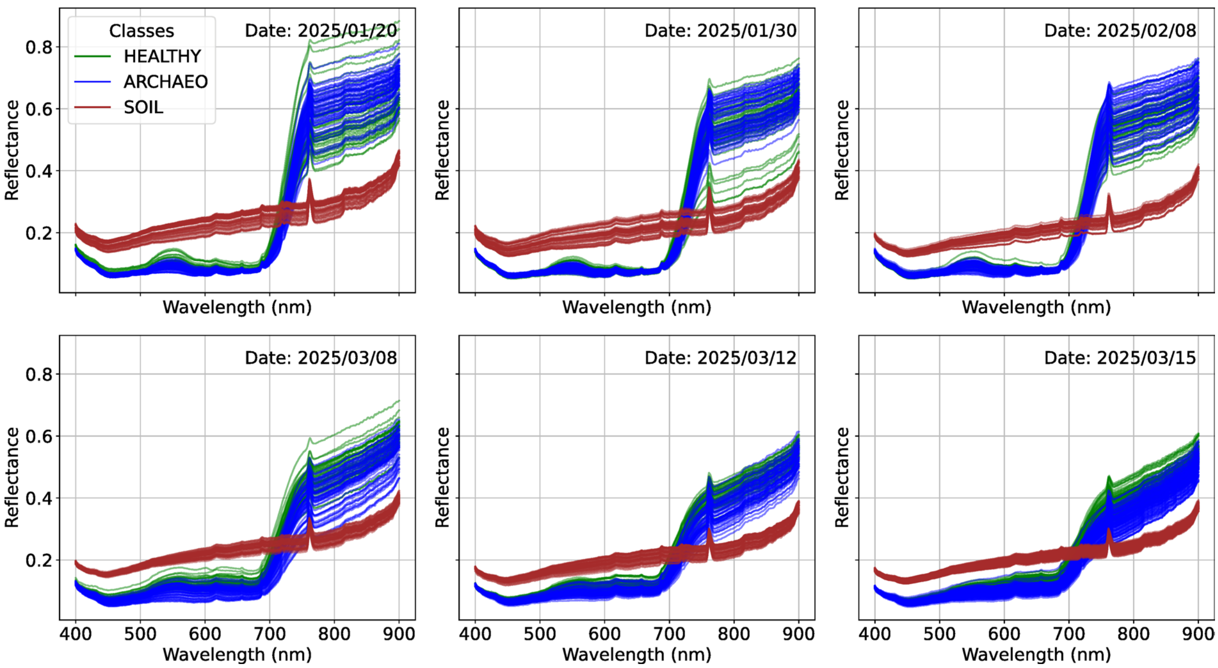

Fig. 1. Measured reflectance signatures in the 2025 campaign.

dates and the numbers of observations in the new campaign are given in Table 1. We note that the spatial coordinates of the individual measurements within the plot were not retained, and the dataset therefore treats all observations as originating from a single spatial unit. The crop signatures over buried remains (cropmarks) will be denoted as ARCHAEO (or A), and every other crop signature will be denoted as HEALTHY (or H). Moreover, a number of signatures over pure soil were acquired. Figure 1 shows the profiles of the reflectance signatures in the wavelength range of 400–900 nm. The new measurements are given in the repository[26] in the file observed_signatures.csv.

The distributions of NDVI and SAVI (with soil-adjustment factor $L = 0.5$) for each date and sample type are shown as boxplots in Fig. 2. The relation between NDVI dynamics and phenological stages for the Alambra test-field has been documented in[27], while broader studies linking vegetation indices to crop phenology are available in e.g.[28–30]. From the data in Fig. 2 we infer that the January–February observations correspond to the peak-greenness period, during which the canopy is fully developed and NDVI values remain high and stable.

In contrast, all March observations fall within the mid- to late-senescence phase, as indicated by the consistent downward shift in both NDVI and SAVI. This decline reflects the post-peak canopy deterioration, with the 15 March measurements representing the late-senescence / pre-harvest stage. It is also worth noting that the cropmark signatures in the present case are associated with somewhat larger NDVI values than the 'healthy' crop signatures. This is a case of positive cropmarks, where the shallow-buried feature creates a moisture- and nutrient-rich microenvironment that supports more vigorous vegetation.

A complementary view is provided by the spectral signatures in Fig. 1, where percentile analysis (not shown) clarifies how these phenological differences manifest across wavelengths. The January–February signature quartiles are narrow across all wavelengths, indicating low within-class variability, with January 20th exhibiting slightly broader tails, suggesting the canopy was still approaching peak greenness. Early-season differences are small but consistent: ARCHAEO absorbs more strongly at 600–700 nm, implying higher chlorophyll than HEALTHY. In March, the quartile envelopes widen sharply at 500–700 nm but tighten in the NIR, indicating that senescence increases heterogeneity mainly through uneven chlorophyll decline and red-edge shifts, while

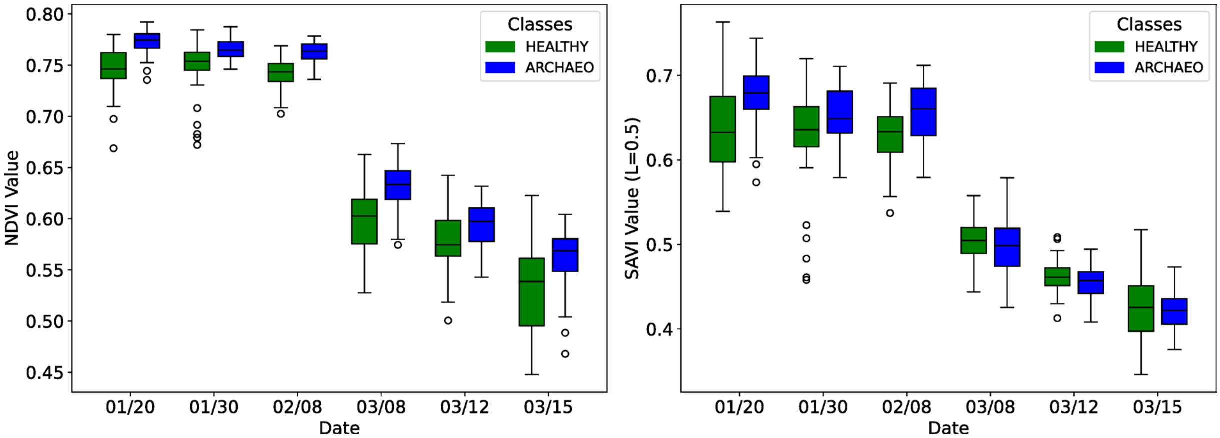

Fig. 2. Boxplots of distributions of the indicated indices per date for measured signatures.

the structural NIR collapse is more uniform. This shift reverses the class ordering: HEALTHY becomes brighter than ARCHAEO across most wavelengths because ARCHAEO senesces earlier and more strongly, producing the VIS/red-edge divergence also reflected in NDVI and SAVI.

Physically-based simulations of signatures

Radiative transfer models encode the physical laws of light transport within vegetation to describe the spectral and directional reflectance of plant canopies, taking into account the geometric and biochemical leaf characteristics, canopy architecture, soil characteristics and the direction of observation. The conceptual and physical basis of radiative transfer modelling has been reviewed e.g. in[31].

PROSAIL[19] is a widely applied canopy reflectance model due to its simplicity and strong overall performance, see e.g.[32]. and references therein. In PROSAIL, canopy is modelled as an infinite turbid medium i.e., a homogeneous and dense layer of very small reflectors (leaves), and radiation transfer is one-dimensional, in the sense that -by horizontal homogeneity- it occurs vertically to canopy thickness[31,33]. For other types of radiative transfer modelling, see e.g.[34]. A thorough comparison of various models see[35,36].

In this work we use PROSAIL within the general context of hybrid inversion[37,40,32,38], in the sense we use synthetic data and machine learning methods to retrieve information about crops. On the other hand, unlike typical applications of the hybrid method, our primary objective is the detection of cropmarks and their classification, rather than the retrieval of specific leaf and canopy parameters. Following our previous work[21] we adopt a pragmatic approach, using PROSAIL to reduce the degrees of freedom of the problem in a physically plausible way, aiming at constructing (physically-based) simulations of a given dataset of observed reflectance (hyperspectral) signatures. The well-known multicollinearity, which is present in the hyperspectral data, see e.g.[41,39]., is primarily due to the physical controls of radiation transfer through vegetation. Hence, the physically-based transcription of observed signatures data is a most reliable means towards synthesizing reflectance signatures with similar physical characteristics, even in the absence of ground-truth physical parameter measurements.

PROSAIL inversion

The PROSAIL physical parameter values and ranges used in the inversion process are generic for grain crops[23] and they are given in Table 2. The nine estimated quantities indicated in bold are varied in the given intervals and they form the group of the physical parameters on which the simulation procedure is based. The values of certain fixed parameters and the choice of the estimated parameters rest on their expected importance, see e.g.[19,42]. and numerical experimentation aiming at plausible distributions of estimated values, as is also discussed below.

In the present work, PROSAIL inversion is performed by constraint-based regularization. We minimize the objective function

$$\text{objective} = \sum (\text{residual})^2 + \text{reg\_weight} \times \sum \left(\frac{p - \text{mid } p}{\text{max}p - \text{mid } p} \right)^2 \tag{1}$$

which is a sum of the square error (difference between the observed and estimated reflectance) plus a term that measures the square deviation of each estimated parameter p from the midpoint of its range in Table 2 scaled by its half-range. The regularization weight (reg_weight) is a pure number, and it is tuned at first by visually assessing plausibility of the distributions of estimated parameter values, fixing reg_weight in the range 0.025 to 0.030. For small values of reg_weight, many parameter histograms appear multimodal and/or concentrated at parameter range boundaries, while for larger values are concentrated too close to the midpoint of the chosen bounds. We shall work with the value reg_weight = 0.030. Minimization of the objective function is done by the 'L-BFGS-B' optimization algorithm of the scipy library of Python 3.9.

Parameter	Value
Leaf structure index, N [-]	1.5
Leaf chlorophyll content, Cab [$\mu g\ cm^{-2}$]	0–80
Leaf carotenoid content, Car [$\mu g\ cm^{-2}$]	0–30
Leaf dry matter content, Cm [$g\ cm^{-2}$]	0.001–0.006
Equivalent water thickness, Cw [cm]	0.004–0.04
Brown pigment content, Cbrown [-]	0–1.0
Leaf anthocyanin content, ant [$\mu g\ cm^{-2}$]	0.0–8
Leaf area index, LAI [$m^2\ m^{-2}$]	0.5–8
Average leaf angle, lidfa [deg]	30–89
Soil reflectance, psoil [-]	variable
Hot-spot size parameter, hspot [$m\ m^{-1}$]	0.05
Sun zenith angle [deg]	0
View zenith angle [deg]	0
Relative azimuth angle [deg]	0

Table 2. The values and the ranges of the PROSAIL input parameters used in this study.

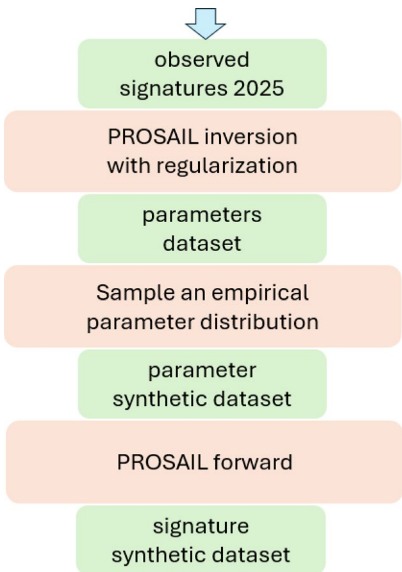

Fig. 3. Flow of work for the simulated signatures production per observation date.

In detail, we proceed as follows. For each date, we use the SOIL measured signatures to estimate a range of psoil parameter by the modelled soil signature used in PROSAIL: psoil*soil_1+(1.0-psoil)*soil_2, where soil_1 and soil_2 correspond to dry and wet soil spectra, respectively. The obtained ranges are used as an a priori range for the psoil parameter for the given dates during the inversion process. Then, for each date and sample type (HEALTHY and ARCHAEO) the PROSAIL curves are fitted to the measured ones by minimizing the objective function obtaining estimates of the nine physical parameters. The estimated values of the physical parameters are given in the repository[26] in the file organized_data.csv. This data is used next in synthetic signatures generation step.

Synthetic dataset construction

Given the dataset of estimated parameters, we derive the corresponding marginal distributions and correlation matrices for each date, which serve as the basis for constructing synthetic spectral signatures. The basic procedure has been described elsewhere[24] hence we shall give a brief account. The difference with[24] is that the 2025 observed dataset is large enough so that we produce synthetic datasets with time-reference. The flow of work is given in Fig. 3.

The PROSAIL inversion and parameter dataset production step has been already discussed in the previous section. Next, for each observation date, we sample from an empirical distribution as defined by the parameter dataset to generate synthetic parameter values which align with the distribution of the estimated values. The necessary input is the correlation matrix and the estimated values dataset of the given date. Then, simulated values are generated as follows: (1) The correlation matrix is Cholesky decomposed. (2) Generate the desired

Description	Dataset
Observed 2012	Observed signatures in Alampra test field in 2012
Observed 2025	Observed signatures in Alampra test field in 2025
Synthetic 2025	Synthetic signatures built from Observed 2025

Table 3. Short-hand description of datasets used in the text.

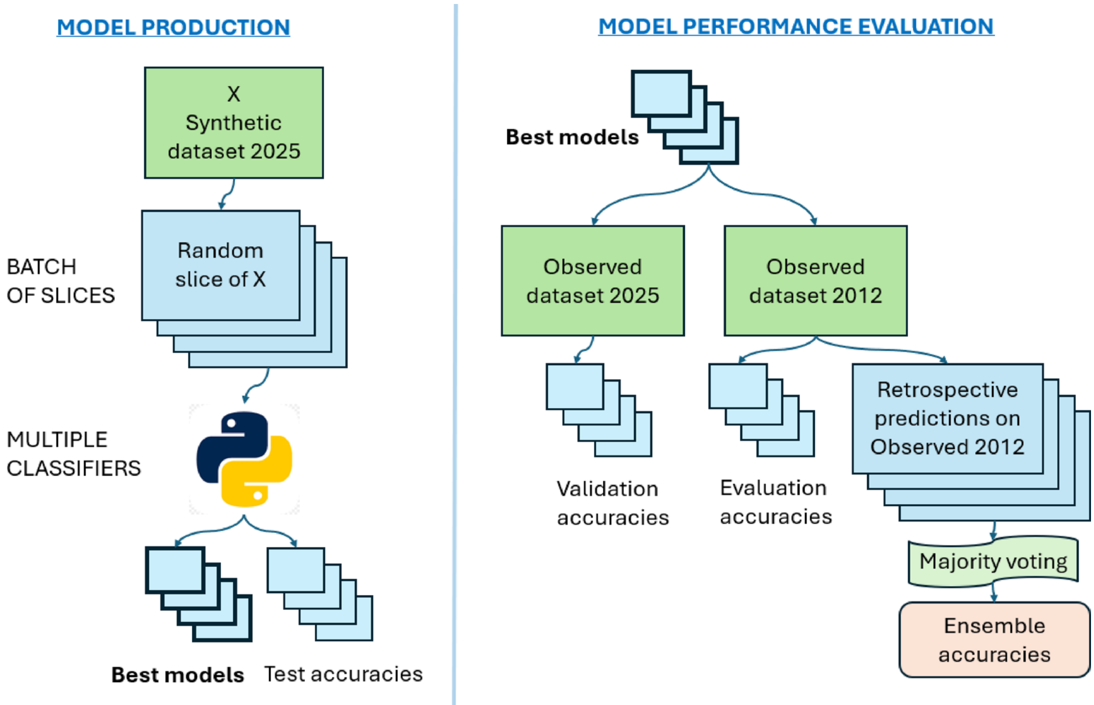

Fig. 4. Chart of machine learning modelling and assessment workflow.

number of normal samples of the 18 physical parameters (9 parameters for each sample type) and apply the Cholesky matrix transformation to produce a multivariate dataset with the desired correlations. (3) Convert the generated multivariate data into a uniform distribution via rank-based transformation. (4). Each uniform variable is mapped to values observed in the estimated parameters dataset using linear interpolation. This ensures that the simulated values retain the main statistical properties of the estimated values dataset. It should be noted that the rank-based mapping restores the empirical marginals, so the final synthetic parameters are non-Gaussian. On the other hand, their correlations remain linear. This is an intentional simplification, as our goal is to assess model behavior using synthetic data that capture only the dominant linear correlation structure of the estimated parameters. Finally, for each date, we run PROSAIL forward, to obtain simulated reflectance signatures. By construction, each synthetic signature is labelled according to date and sample type (HEALTHY, ARCHAEO). In repository[26] a simulator script is provided to generate the desired number of signature simulations.

Machine learning modelling

With the synthetic signatures at hand, one may model this dataset, by any method of one's preference, and subsequently use this model as a detection algorithm for identifying the A signatures in another observed dataset. In the present work, we use the measured signatures of the 2012 campaign. For ease of reference, we summarize in Table 3 the short-hand that we will employ in describing the datasets used in the present work.

A set machine learning models are trained on the Synthetic 2025 and their performance in identifying the ARCHAEO/HEALTHY (A/H) signatures in the Observed 2012 dataset is evaluated. Our goal is to evaluate how effectively the synthetic dataset, derived from observations at a specific point in time, can serve as the foundation for retrospective predictive modeling frameworks in cropmark detection.

Figure 4 presents the workflow of the machine learning modelling and assessment. We use the generating script of[42] to produce a set X of 50,000 synthetic signatures from each sample type (A and H) and date, that is, totally 600,000 signatures. In what follows, X represents the Synthetic 2025 dataset. The dataset size is an important control parameter of the machine learning predictive modelling performance; hence, we train models on *slices* of the Synthetic 2025 dataset, that is, a sample of signatures randomly selected from X without replacement. To achieve stability of the results, the models are trained on *batches* of slices of the same size. This approach allows us to perform comparative robustness analysis, systematically examining how model performance adapts to

Parameter	Value
Batch size	200
Slice sizes	50, 75, 100, 125, 150, 200, 250
Noise levels	0–5% (increments of 0.5)

Table 4. Training control parameters.

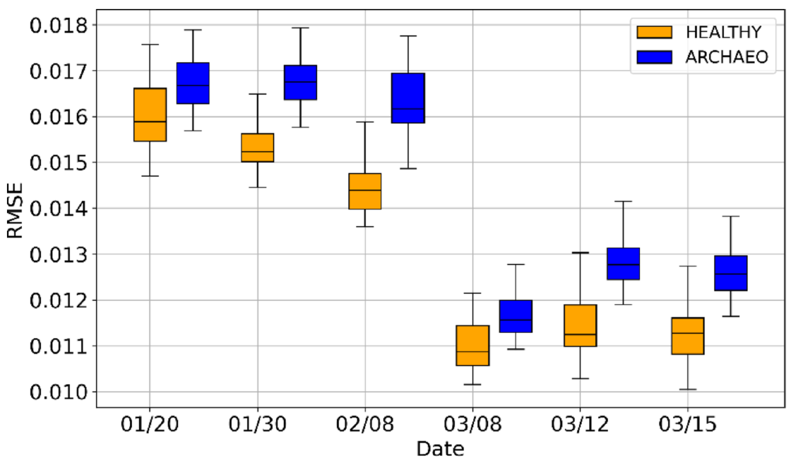

Fig. 5. Box-plots of the distributions of root-mean-square errors of estimated reflectance signatures for the different dates and sample types.

varying dataset sizes and investigating how patterns observed in smaller subsets remain stable when scaling up the synthetic dataset. We also examine the effect of injecting noise into the synthetic signatures. Input noise has long been used as a regularization technique for reducing generalization error in machine learning[43,44]. Input perturbations reduce overfitting by making models less reliant on perfectly clean synthetic signatures and more tolerant of real-world variability. In our experiments, we add white noise of up to 5% to each wavelength value of every synthetic signature. This noise level was selected based on the usual radiometric uncertainties of most satellite optical sensors that have been explored in the past in the context of remote sensing archaeological prospection campaigns. The batch size, slice sizes and noise levels employed in our analysis are presented in Table 4. Slice size refers to the number of each type of signature, A and H, for each date.

For the modelling stage, we selected a diverse set of classifiers to capture different learning paradigms and inductive biases, drawing on established evidence regarding their performance, scalability, robustness, and interpretability[45–48]. The set includes linear models (Logistic Regression, Ridge), probabilistic methods (Gaussian Naive Bayes), instance-based learning (K-Nearest Neighbors), tree-based ensembles (Extra Trees, Random Forests), gradient boosting (XGBoost), kernel-based methods (SVC), and neural networks (MLPs), thereby spanning a broad methodological spectrum from simple linear decision rules to highly flexible non-linear learners. All classifiers were implemented in Python 3.9 using scikit-learn (v1.6.1) and XGBoost (v2.1.4), and evaluated through GridSearchCV under default hyperparameter settings, with only minimal adjustments where necessary.

Each classifier is employed via grid-search in training multiple models across different slices within every batch. We use 80 – 20 split for training and testing and 5-fold cross validation (cv = 5). The performance of each model is evaluated using (mean cross-validation) accuracy as the primary metric. For each slice in every batch, the model with the highest accuracy is selected as the optimal choice (best model).

Results

Prosail fit and statistics of the estimated parameters dataset

Boxplots of the root-mean-square error (RMSE) between the estimated and the measured reflectance signatures are shown in Fig. 5. We observe that the overall RMSE of estimated reflectance is roughly 1.5%. We may also note the somewhat larger error values in the first three dates of measurements. These results illustrate the quality of the observed signatures' fit by PROSAIL, and hence quantify the uncertainty of the inversion process at the level of signatures.

Across the January–February dates, the parameter distributions show patterns that are broadly consistent with a peak-greenness canopy inferred from NDVI distributions (Fig. 2). Chlorophyll content (Cab) is relatively high for both sample types during this period, with ARCHAEO exhibiting consistently higher quartiles than HEALTHY, in line with the slightly higher NDVI and SAVI values observed earlier in the season. Carotenoid content (Car) remains stable and within a narrow range, as expected for a fully green canopy where protective pigments are present but not yet influenced by senescence. Brown-pigment absorption (Cbrown) is low during these dates, reflecting minimal accumulation of senescence-related compounds. Leaf mass per area (Cm) and leaf water content (Cw) show only small variations, consistent with structurally intact, hydrated foliage. Leaf angle distribution (lidfa) and LAI also fall within ranges typical of a dense, functioning canopy.

The March dates exhibit parameter shifts that align with mid- to late-senescence, as indicated by the NDVI and SAVI decline (Fig. 2). Cab shows a pronounced decrease across all March observations, reflecting the expected reduction in chlorophyll during senescence. In contrast, Car remains comparatively stable, producing a widening separation between chlorophyll and carotenoid levels that is typical of progressing senescence, where chlorophyll declines more rapidly than accessory pigments. Cbrown increases sharply for both sample types, indicating the accumulation of senescence-related absorbing compounds associated with pigment breakdown. Cm and Cw remain relatively stable, with only subtle reductions, which is reasonable for this stage where pigment loss precedes major structural or water-content changes. Leaf angle (lidfa) increases through March, suggesting a shift toward more open canopies as leaves age. LAI also declines relative to earlier dates, consistent with canopy thinning. These parameter trajectories do not determine the phenological staging but support the interpretation that all March observations fall within mid- to late-senescence, with March 15th representing the most advanced point within this phase, showing the strongest pigment degradation and highest Cbrown levels, consistent with a pre-harvest condition. On all dates, the psoil parameter mirrors the fitted soil spectra, with a general downward trend consistent with increasing soil moisture, except for March 8th, which reflects the measured soil signatures for that date.

The correlations between the physical parameters corresponding to different observation dates are illustrated in Fig. 6 in the form of heatmap correlations matrices. A subscript A for ARCHAEO and H for HEALTHY, indicates the class where physical parameter belongs to. Given that we estimate 9 physical parameters, these matrices are 18×18. Correlations between the parameter values of the same or different sample types further illustrate the interpretations of the observed signatures in terms of leaf and canopy parameters. At the same time, they are a necessary piece of information for the simulation generation part of our analysis.

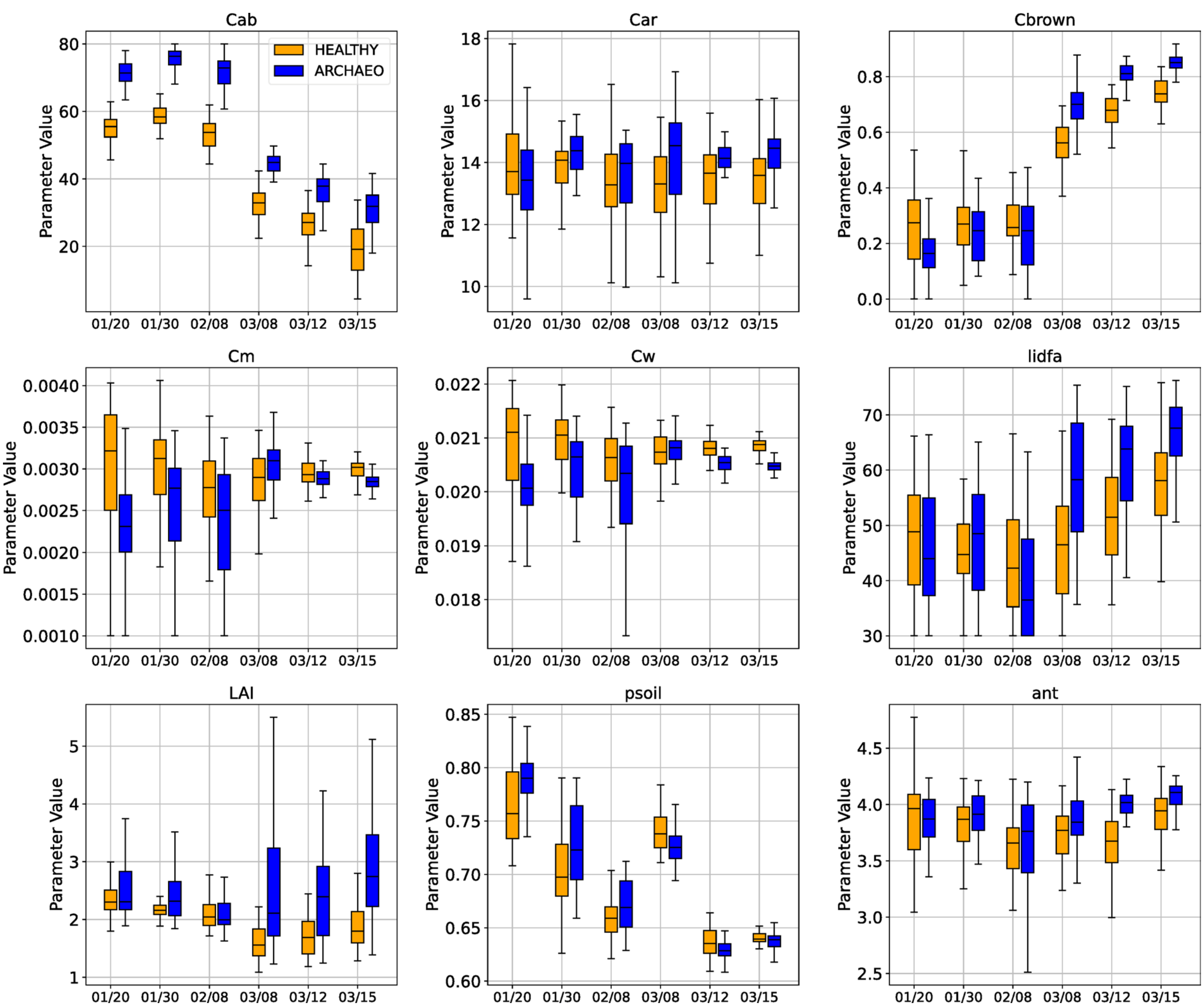

Fig. 6. Box-plots of the distributions of the estimated physical parameters for the different dates and sample types.

A consistent feature of the correlation matrices (Fig. 6) is the strong association between Cm and Cw, and between LAI and lidfa, within each sample type. The psoil parameter is negatively correlated with nearly all other variables across dates. Carotenoids show a moderate correlation with chlorophyll in the first two dates, which weakens substantially thereafter. Qualitatively, the matrices are close to block-diagonal, though to varying degrees across dates. As discussed later, these structural differences influence both the character of the synthetic datasets and the resulting classification performance.

Quality of the synthetic dataset

Visual inspection and statistical tests indicate that the simulated parameter distributions closely match the estimated ones across all dates and sample types. KS tests yield p-values of 1.0, Mann–Whitney U tests exceed 0.90, and chi-square tests also return p-values of 1.0, all pointing to indistinguishable marginal distributions. Wasserstein distances remain near zero, and KL divergence values are mostly negligible, with only occasional moderate deviations. Overall, the synthetic data successfully reproduces the marginal behavior of the original parameters.

The correlation structure is likewise well preserved. Simulated correlation matrices visually resemble the estimated ones, with differences limited to small quantitative shifts. Matrix-level comparisons confirm this: Frobenius distances of 0.66–1.13 and eigenvalue distances of 0.42–0.84 indicate modest deviations in global structure and variance composition. Although the first three dates show slightly larger discrepancies, consistent with their higher RMSE values (Fig. 1), the overall dependency patterns remain stable across observations.

Next, we assess whether the simulated signatures remain faithful to the manifold of real reflectance signatures. Due to the difference of observed and synthetic datasets size, and the lack of any natural pairing between the synthetic and observed signatures, we computed the Spectral Angle Mapper (SAM) between each synthetic signature and its nearest neighbor in the observed dataset. Hence, we obtained a distribution of nearest-neighbor SAM (n-n SAM) per observation date and signature type. The corresponding boxplots are shown in Fig. 7a.

Across all dates and for both classes, SAM values remain consistently small. Median angles range from approximately 0.029 to 0.043 radians ($\approx 1.7°–2.5°$), with narrow interquartile ranges across all datasets, indicating tight clustering of synthetic spectra around their real counterparts. Even the upper tail of the distributions remains modest: the 95th percentiles fall between 0.042 and 0.079 radians for the H class and between 0.049 and 0.077 radians for the type A class. These values show that even the most distant synthetic spectra remain close to real observations.

The regularization weight (reg_weight) is the parameter used during PROSAIL inversion to estimate biophysical parameters from the observed reflectance signatures. We now examine its influence on the quality of the synthetic signature dataset, as quantified by the nearest-neighbor spectral angle (n-n SAM). Figure 7b shows a weak upward trend in n-n SAM with increasing reg_weight, consistent across all dates and sample types. This indicates that within the explored range and particularly for values below 0.03, the choice of reg_weight has only a minor effect on the fidelity of the synthetic signatures. On the basis of this observation, we have employed a plausibility argument on physical parameter distributions to narrow down the values of interest between 0.025 and 0.030, as mentioned previously. It is also worth emphasizing that the issue of the quality of

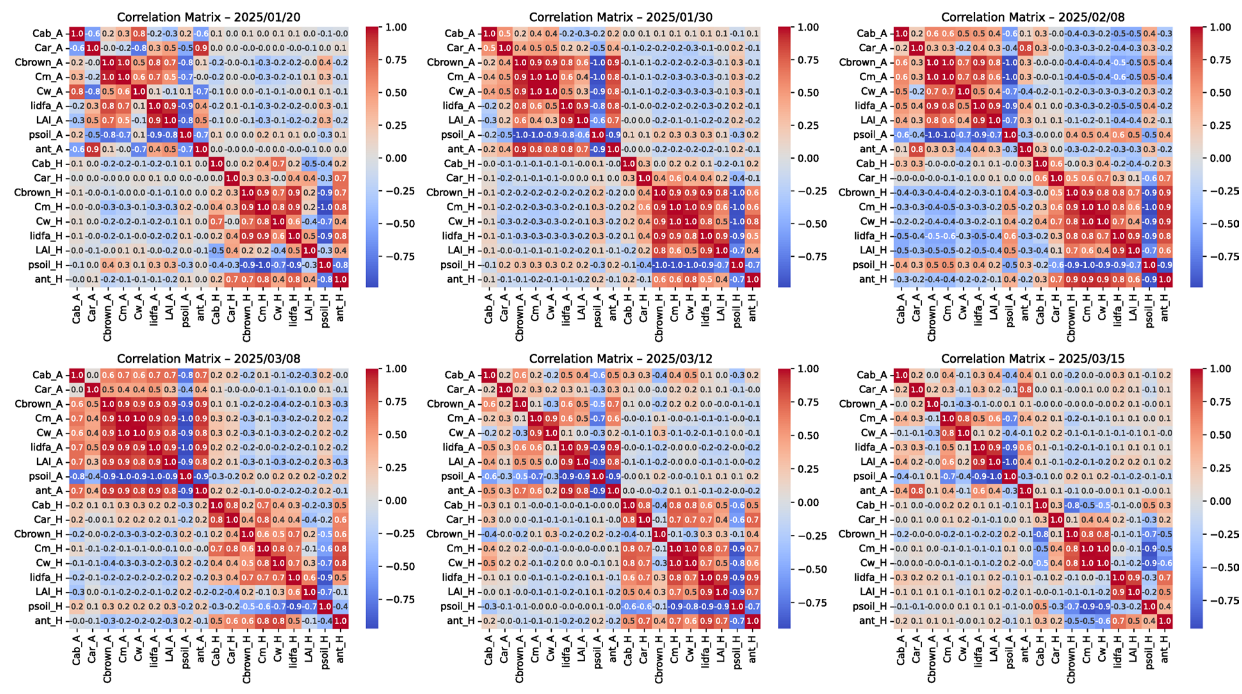

Fig. 7. Correlation matrices of the estimated physical parameters for the different dates

the synthetic signatures dataset is explored in greater depth below in the context of machine-learning modelling of the synthetic signatures, assessing the ability of the models to recognize the original (measured) signatures.

Machine learning modelling assessment results

Classifier dominance presented in Fig. 8 follows a clear seasonal structure that mirrors the phenological progression inferred from the PROSAIL parameters. During the peak-greenness period (January and February), the spectral space is biochemically simple and highly separable, with high Cab, low Cbrown, saturated NDVI/SAVI, and limited structural heterogeneity. Under these conditions, linear decision boundaries are strongly favored: the RidgeClassifier overwhelmingly dominates, followed by ExtraTrees, while nonlinear or distance-based models contribute minimally. This reflects the fact that reflectance during peak greenness is governed by smooth, monotonic pigment–NIR relationships that are well captured by linear models. ExtraTrees and RandomForest appear in secondary roles, consistent with their ability to capture modest interaction effects even in relatively simple feature spaces. The stability of this ranking under both noise levels further indicates that separability during peak greenness is robust and largely insensitive to perturbations.

A marked shift occurs across all mid- to late-senescence dates in March, with March 8th simply representing the earliest observation within this phase. During this period, the classifier landscape changes substantially: the MLP becomes the dominant model, followed by ExtraTrees and RandomForest, while RidgeClassifier drops sharply in importance. This reflects the increased spectral and structural heterogeneity characteristic of senescence, as Cab declines, Cbrown rises, lidfa becomes more dispersed, and NDVI/SAVI distributions broaden. A minor peculiarity of the March 8th data is that psoil does not follow the decreasing trend seen on later dates, a consequence of unexpectedly high measured soil reflectances; this likely influenced the estimation of some parameters and contributed to the modelling behavior on that date. By March 12th and 15th, the dominance patterns strengthen further, and the difference between these two dates becomes apparent even at the classifier-aggregation level: the "late-season" dominance scores show a clear increase in MLP weight and a redistribution among the secondary models, indicating that March 15th corresponds to a more advanced stage within the senescence window than March 12th. This progression is consistent with a canopy that becomes increasingly heterogeneous as senescence advances, with greater structural dispersion and more variable background contributions. The resulting fragmentation of the spectral landscape naturally favors flexible, nonlinear decision surfaces, explaining the pronounced rise of the MLP and the shifting roles of the tree-based models.

Model assessment proceeds as follows. For each "best model," the test accuracy on the remaining 20% of the split is recorded during training (Fig. 4). Each best model is then applied to the Observed 2025 dataset, which was held out as an additional test set, to assess how well a model trained on the Synthetic 2025 slices performs on the original signatures. We refer to these as validation accuracies. Finally, each best model is evaluated on the past Observed 2012 dataset in two ways: (i) by the directly obtained accuracies (evaluation accuracies) and (ii) by applying a majority-vote rule across all models within each slice-size batch, yielding ensemble accuracies (Fig. 4). Corresponding ensemble-level recall and precision scores are also computed, representing the predictive performance for each slice size and date.

The machine-learning performance results are shown in Fig. 9. Each subfigure is labelled by date and noise condition (5%). Peak-greenness dates (January–February) appear on the left of the multiplot, and mid- to late-senescence dates (March) on the right. Test, validation, and evaluation accuracies are presented as violin plots, which illustrate their distributions across models for each slice size. Median values are indicated by single-dash markers. Ensemble accuracies, which are single values per slice size, are plotted as line-connected cross markers.

We observe varied patterns in the test (blue) and internal (green) accuracies across the different dates. Using the median accuracy as the metric, in the peak greenness period, test accuracies are consistently high and tend to increase with the slice size. The first and the last March dates exhibit somewhat lower test accuracies, while March 12th returns lower accuracy. This indicates that the algorithm captures signature variability differently across dates for the given slice-size range. Validation accuracies (green), defined as performance on the Observed

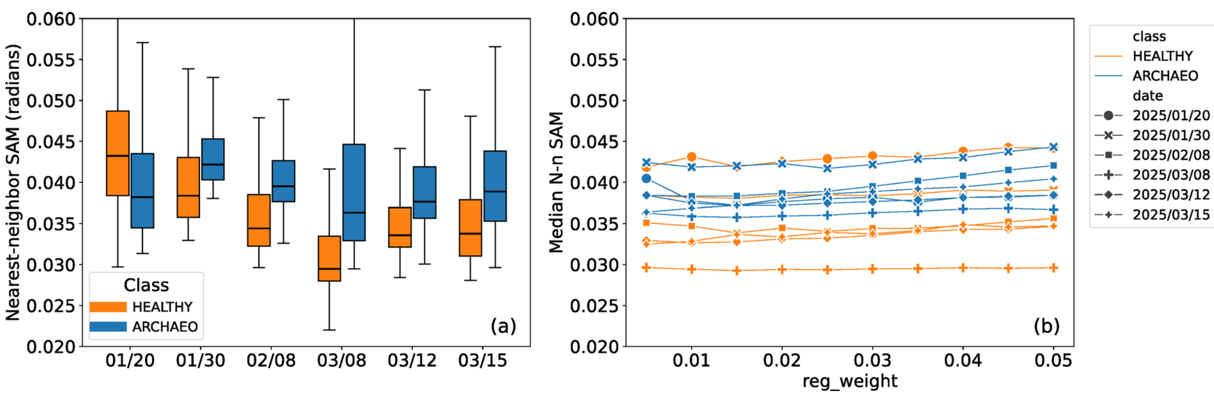

Fig. 8. (**a**) Boxplots of nearest-neighbor spectral angle between observed and synthetic reflectance signatures. (**b**) Median values of the n-n SAM against the regularization weight.

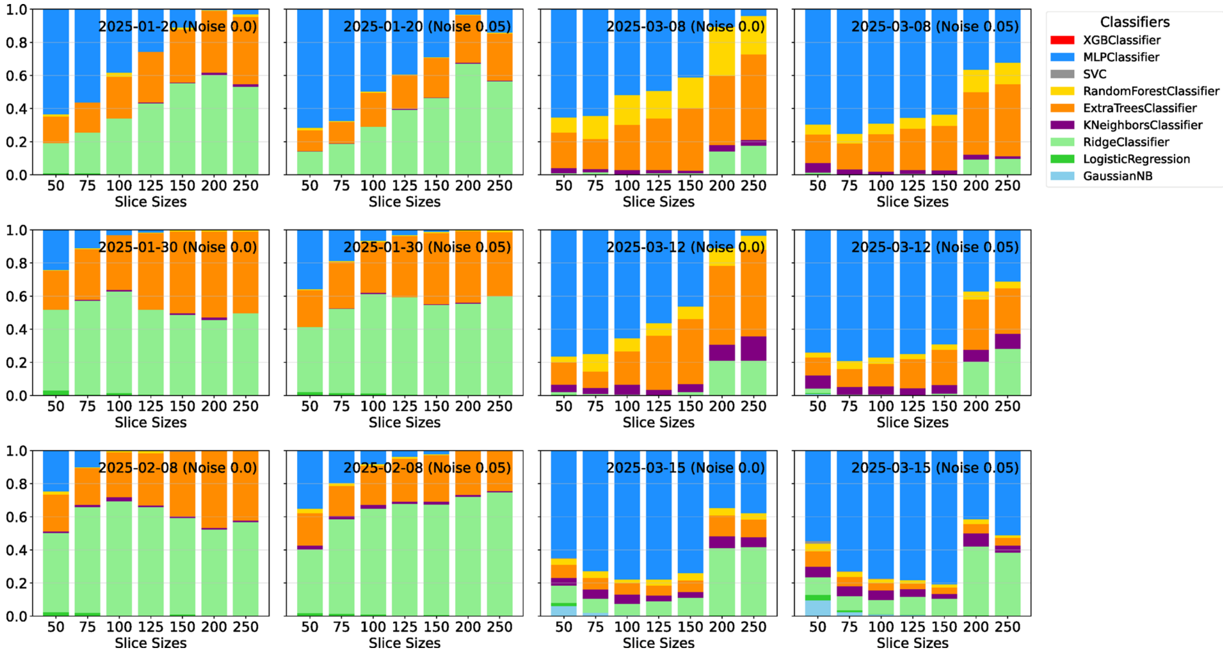

Fig. 9. Proportion of occurrences of each type of classifier as optimal choice for different slice sizes and dates..

2025 dataset, show a nearly uniform and slightly decreasing trend around 90%, except on March 12th where they fall below 80%. That means that for most of the observation dates, the models 'recognize' the original measured signatures at a rate near 90%. Perhaps the most consistent behavior from the validation accuracies perspective is exhibited in the February 8th and March 8th observation dates. We may also note that the previous observations hold with or without added noise.

External (red) and ensemble (cross) classification scores quantify the performance of the models on the Observed 2012 dataset, that is, the rate the models 'recognize' reflectance signatures observed in the past (thirteen years earlier). The violin plots show large spreads, so we focus on the ensemble scores. These represent the success rate of the ensemble of best classifiers, whose predictions are aggregated by majority voting for each slice size and date.

Ensemble performance during the peak-greenness period (January-February) is characterized by high accuracy (≈ 0.85–0.88), perfect precision (1.000), and moderately high recall (≈ 0.70–0.76). These metrics remain stable across slice sizes and noise levels, reflecting the biochemical and structural uniformity of the canopy during peak greenness. Slice size controls only how densely the synthetic reflectance manifold is sampled during training. The flat response across slice sizes therefore indicates that the manifold is smooth, low-curvature, and well-separated, requiring little sampling density to stabilize predictions. Even at 30 January, where some slice sizes show notable accuracy and recall decline under noise-free conditions, the addition of noise (0.05) regularizes the ensemble and restores stable performance. Overall, the peak-greenness period represents a high-confidence, high-stability prediction regime in which the underlying spectral structure is simple and easily learned.

A distinct shift occurs across all mid- to late-senescence dates (March) where ensemble performance becomes strongly slice-size dependent. Small to mid-range slices (50–125) achieve the highest accuracies (≈ 0.89–0.93) and strong recall (≈ 0.77–0.85), while larger slices begin to degrade performance. Noise injection provides a modest improvement. This behavior reflects the increased irregularity and heterogeneity of the reflectance manifold during senescence, shaped by declining Cab, rising Cbrown, more dispersed lidfa, and broader NDVI/SAVI distributions. The March 8th data also reflect parameter influences linked to the atypical soil reflectances measured on that date. In this regime, the manifold becomes locally noisy and less separable, and classifiers respond differently depending on how much of this complexity they encounter during training.

By the later senescence dates (March 12th and 15th), ensemble behavior becomes highly fragmented: small slices yield high precision but extremely low recall (e.g., recall ≈ 0.09–0.30), indicating overconfident predictions that miss most ARCHAEO signatures, while mid-range slices (≈ 150) provide the best balance of accuracy and recall. The slightly stronger deterioration observed on March 15th is consistent with it being the most advanced point within the senescence window, approaching pre-harvest. Larger slices again degrade performance because dense sampling forces the ensemble to confront the full spectral heterogeneity of advanced senescence. These patterns mirror the PROSAIL-derived biochemical and structural changes (Fig. 10), showing that as the canopy becomes spectrally complex, prediction quality becomes tightly linked to how densely the reflectance manifold is sampled. Across all dates and slice sizes, noise injection has a consistently non-negative effect on model performance.

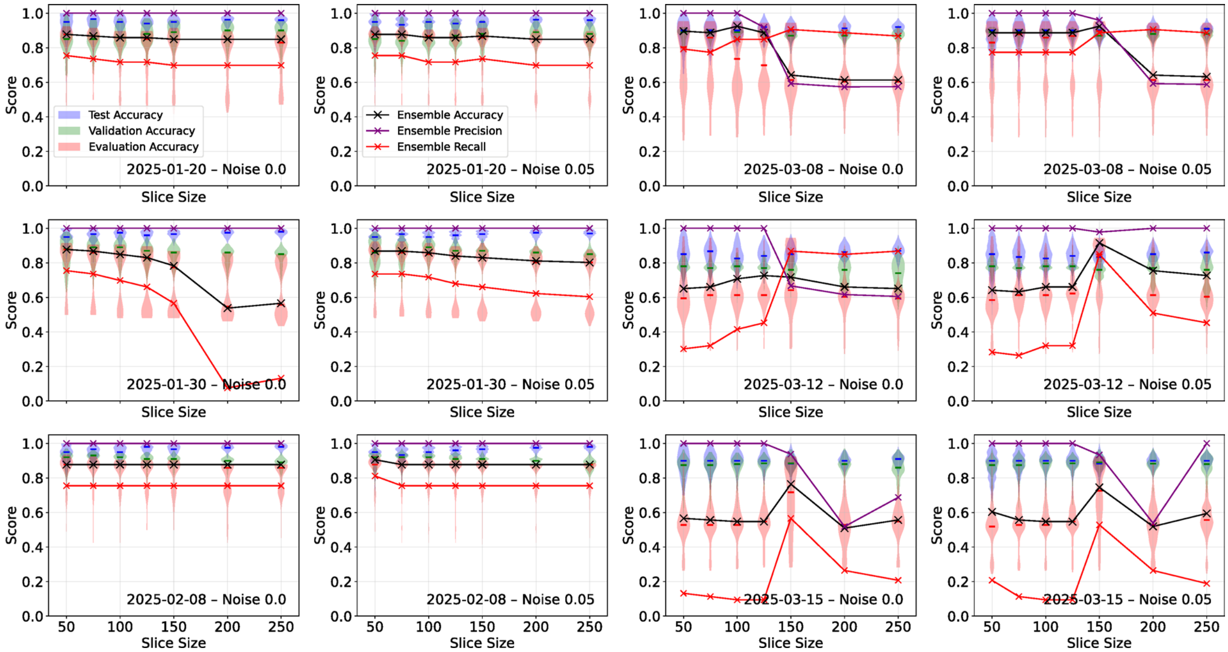

Fig. 10. Test, validation and evaluation accuracy violin plots and majority voting classification scores for different slice sizes and dates.

Summary and discussion

This study explored the feasibility of retrospective cropmark modelling by constructing synthetic cropmark and healthy crop reflectance signatures from observations collected at the Alampra, Cyprus test field during the January–March 2025 campaign. We used PROSAIL inversion with constrained regularization and statistically generated parameter ensembles to produce synthetic signatures. These signatures closely reproduce the distributions and linear correlations of the estimated biophysical parameters. A diverse set of machine-learning classifiers was then trained on these synthetic datasets and evaluated on past observations from the same field (2012). Ensemble classification scores were used as performance metrics.

Ensemble precision and recall reveal clear date-dependent patterns. Peak-greenness signatures (January–early February) yield the most reliable models, with consistently high accuracy, stable recall, and very high precision. In contrast, the mid- to late-senescence period in March shows more variable behavior. Earliest March data still perform reasonably well, though gains in recall at larger slice sizes tend to come with reduced precision. As senescence progresses through March, ensemble behavior becomes increasingly irregular, with recall showing marked instability and precision–recall trade-offs becoming more pronounced. This gradual deterioration is consistent with the canopy becoming more heterogeneous as senescence advances. Across all dates, noise injection generally has a non-negative effect, occasionally acting as a mild regularizer. Slice size 150 marks a broad transition point beyond which performance tends to deteriorate on the more unstable, senescence-stage datasets.

The entire modelling work of our ensemble classifier rests on the physically based construction of the synthetic reflectance signatures. Our approach combines PROSAIL inversion with statistically generated parameter ensembles to produce synthetic signatures that preserve the structure and variability of the observed spectra. The inversion step serves primarily as a physically informed reduction of dimensionality, ensuring that the simulated signatures are built upon reasonable, though not necessarily true, biophysical parameter values. This physically grounded construction is central to the framework. It provides the dataset on which the ensemble classifier is trained and underpins its ability to perform retrospective detection by anchoring the synthetic data in the spectral behavior of the test-field conditions.

While the framework provides a coherent basis for retrospective detection, it is subject to several constraints that limit the generality of the findings. This study is based on observations from a single controlled test-plot, and therefore does not capture the broader variability of cropmarks across soil types, hydrological regimes, crop varieties, and management practices. The synthetic-signature workflow also depends on PROSAIL inversion and statistical modelling choices that may behave differently under more heterogeneous or stressed conditions. Model performance is further influenced by phenological stage, as evidenced by the strong date-dependent patterns observed in the ensemble scores. Finally, the retrospective evaluation is limited to a single past dataset, and thus does not yet establish how robust the approach would be across multiple years or under different sensor characteristics.

Future work will focus on extending the framework beyond the controlled conditions of the present study and addressing the broader issue of scalability. A priority is the acquisition of new hyperspectral signatures across multiple fields, soil types, hydrological regimes, and crop varieties, supported by detailed ground observations of phenology, canopy structure, and environmental conditions. Such datasets would allow the synthetic-signature workflow to be calibrated and validated under the heterogeneity of real archaeological landscapes, strengthening both its physical realism and external validity. Lessons learned from these expanded studies would also contribute to the way for retrospective predictive modelling at landscape scale. In particular, enriched spectral libraries could enable the systematic exploitation of archived aerial and satellite imagery, offering the possibility of recovering archaeological proxies that may no longer be detectable today due to natural or anthropogenic change. Synthetic datasets constructed from these new libraries, augmented by accumulated expert knowledge in both physical modelling and machine-learning design, will be central to advancing this capability.

Data availability
The datasets generated and analysed during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.15767205.

Received: 13 July 2025; Accepted: 18 March 2026
Published online: 26 March 2026

References
1. Giardino, M. J. A history of NASA remote sensing contributions to archaeology. *J. Archaeol. Sci.* **38**(9), 2003–2009 (2011).
2. Forte, M. & Campana, S. Digital methods and remote sensing in archaeology. *Archaeol. Age Sens.* (2016).
3. Lasaponara, R. & Masini, N. Satellite remote sensing in archaeology: Past, present and future perspectives. *J. Archaeol. Sci.* **38**(9), 1995–2002 (2011).
4. Moriarty, C., Cowley, D. C., Wade, T. & Nichol, C. J. Deploying multispectral remote sensing for multi-temporal analysis of archaeological crop stress at Ravenshall, Fife, Scotland. *Archaeol. Prospection*. **26**(1), 33–46 (2019).
5. Agapiou, A., Hegyi, A. & Stavilă, A. Observations of Archaeological Proxies through Phenological Analysis over the Megafort of Csanádpalota-Juhász T. tanya in Hungary Using Sentinel-2 Images. *Remote Sens.* **15**, 464 (2023).
6. Czajlik, Z. et al. Cropmarks in aerial archaeology: New lessons from an old story. *Remote Sens.* **13**(6), 1126 (2021).
7. Stewart, C., Labrèche, G. & González, D. L. A pilot study on remote sensing and citizen science for archaeological prospection. *Remote Sens.* **12**(17), 2795 (2020).
8. Cowley, D. C. Creating the cropmark archaeological record in East Lothian, southeast Scotland. Prehistory without Borders: Prehistoric Archaeology of the Tyne-Forth Region. Oxford: Oxbow, 59–70 (2016).
9. Banaszek, Ł., Cowley, D. C. & Middleton, M. Towards national archaeological mapping. Assessing source data and methodology—A case study from Scotland. *Geosciences* **8**(8), 272 (2018).
10. Campana, S. 'Total archaeology' to reduce the need for rescue archaeology: The BREBEMI Project (Italy). *Remote Sensing for Archaeological Heritage Management* 33–41 (2011).
11. Agapiou, A., Lysandrou, V., Sarris, A., Papadopoulos, N. & Hadjimitsis, D. G. Fusion of satellite multispectral images based on ground-penetrating radar (GPR) data for the investigation of buried concealed archaeological remains. *Geosciences* **7**(2), 40 (2017).
12. Zhang, X. et al. Multimodal remote sensing image registration methods and advancements: A survey. *Remote Sens.* **13**(24), 5128 (2021).
13. Menze, B. H. & Ur, J. A. Multitemporal fusion for the detection of static spatial patterns in multispectral satellite images—With application to archaeological survey. *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.* **7**(8), 3513–3524 (2014).
14. Opitz, R. & Herrmann, J. Recent trends and long-standing problems in archaeological remote sensing. *J. Comput. Appl. Archaeol.* **1**(1), 19–41 (2018).
15. Agapiou, A. & Sarris, A. Beyond GIS layering: Challenging the (Re) use and fusion of archaeological prospection data based on Bayesian Neural Networks (BNN). *Remote Sens.* **10**(11), 1762 (2018).
16. Gojda, M. & Hejcman, M. Cropmarks in main field crops enable the identification of a wide spectrum of buried features on archaeological sites in Central Europe. *J. Archaeol. Sci.* **39**(6), 1655–1664 (2012).
17. Hejcman, M. & Smrž, Z. Cropmarks in stands of cereals, legumes and winter rape indicate sub-soil archaeological features in the agricultural landscape of Central Europe. *Agric. Ecosyst. Environ.* **138**(3–4), 348–354 (2010).
18. Menze, B. H. & Ur, J. A. Mapping patterns of long-term settlement in Northern Mesopotamia at a large scale. *Proc. Natl. Acad. Sci. USA* **109**(14), E778–E787 (2012).
19. Jacquemoud, S. et al. PROSPECT+ SAIL models: A review of use for vegetation characterization. *Remote Sens. Environ.* **113**, 56–66 (2009).
20. Agapiou, A. & Gravanis, E. A Machine-Learning-Assisted Classification Algorithm for the Detection of Archaeological Proxies (Cropmarks) Based on Reflectance Signatures. *Remote Sens.* **16**(10), 1705 (2024).
21. Gravanis, E. & Agapiou, A. *Physically-based detection algorithm of buried archaeological remains using spectral signatures* (IEEE Access, 2024).
22. Gravanis, E. & Agapiou, A. Archaeological cropmark synthetic signatures (ACSS). *Mendeley Data*. **V1** https://doi.org/10.17632/465vpcsd84.1 (2025).
23. Gravanis, E. & Agapiou, A. *Descriptor: Archaeological cropmark synthetic signatures (ACSS)* (IEEE Data Descriptions, 2025).
24. Gravanis, E. & Agapiou, A. *Physically-based predictive modelling of archaeological proxies using cropmarks* (2025).
25. Agapiou, A. & Thesis, P. D. *Development of a Novel Methodology for the Detection of Buried Archaeological Remains Using Remote Sensing Techniques* (Cyprus University of Technology, 2013). https://hdl.handle.net/20.500.14279/877 (accessed on 8 May 2024).
26. Gravanis, E. & Agapiou, A. Archaeological cropmark synthetic signatures 2 (ACSS 2) [Dataset]. Zenodo. (2025). https://doi.org/10.5281/zenodo.15767205
27. Agapiou, A., Hadjimitsis, D. G., Sarris, A., Georgopoulos, A. & Alexakis, D. D. Optimum temporal and spectral window for monitoring crop marks over archaeological remains in the Mediterranean region. *J. Archaeol. Sci.* **40**(3), 1479–1492 (2013).
28. Bhatti, M. T., Gilani, H., Ashraf, M., Iqbal, M. S. & Munir, S. Field validation of NDVI to identify crop phenological signatures. *Precision Agric.* **25**(5), 2245–2270 (2024).
29. Pan, Z. et al. Mapping crop phenology using NDVI time-series derived from HJ-1 A/B data. *Int. J. Appl. Earth Obs. Geoinf.* **34**, 188–197 (2015).
30. Zeng, L., Wardlow, B. D., Xiang, D., Hu, S. & Li, D. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. *Remote Sens. Environ.* **237**, 111511 (2020).
31. Myneni, R. B., Ross, J. & Asrar, G. A review on the theory of photon transport in leaf canopies. *Agric. For. Meteorol.* **45**(1–2), 1–153 (1989).
32. Berger, K. et al. Evaluation of the PROSAIL model capabilities for future hyperspectral model environments: A review study. *Remote Sens.* **10**(1), 85 (2018).
33. Verhoef, W., Jia, L., Xiao, Q. & Su, Z. Unified optical-thermal four-stream radiative transfer theory for homogeneous vegetation canopies. *IEEE Trans. Geosci. Remote Sens.* **45**(6), 1808–1822 (2007).

34. Dorigo, W. A. et al. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *Int. J. Appl. Earth Obs. Geoinf.* **9**(2), 165–193 (2007).
35. Pinty, B. et al. Radiation Transfer Model Intercomparison (RAMI) exercise: Results from the second phase. *J. Geophys. Research: Atmos.* **109**, D6 (2004).
36. Widlowski, J. L. et al. The fourth phase of the radiative transfer model intercomparison (RAMI) exercise: Actual canopy scenarios and conformity testing. *Remote Sens. Environ.* **169**, 418–437 (2015).
37. Verrelst, J. et al. Optical remote sensing and the retrieval of terrestrial vegetation bio-geophysical properties–A review. *ISPRS J. Photogrammetry Remote Sens.* **108**, 273–290 (2015).
38. Abdelbaki, A. & Udelhoven, T. A review of hybrid approaches for quantitative assessment of crop traits using optical remote sensing: research trends and future directions. *Remote Sens.* **14**(15), 3515 (2022).
39. Verrelst, J. et al. Quantifying vegetation biophysical variables from imaging spectroscopy data: A review on retrieval methods. *Surv. Geophys.* **40**, 589–629 (2019).
40. Liang, L. et al. Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sens. Environ.* **165**, 123–134 (2015).
41. Atzberger, C., Guérif, M., Baret, F. & Werner, W. Comparative analysis of three chemometric techniques for the spectroradiometric assessment of canopy chlorophyll content in winter wheat. *Comput. Electron. Agric.* **73**(2), 165–173 (2010).
42. Danner, M., Berger, K., Wocher, M., Mauser, W. & Hank, T. Fitted PROSAIL parameterization of leaf inclinations, water content and brown pigment content for winter wheat and maize canopies. *Remote Sens.* **11**(10), 1150 (2019).
43. Sietsma, J. & Dow, R. J. Creating artificial neural networks that generalize. *Neural Netw.* **4**(1), 67–79 (1991).
44. Bishop, C. M. & Nasrabadi, N. M. *Pattern recognition and machine learning. Vol. 4. No. 4* (Springer, 2006).
45. Sen, P. C., Hajra, M. & Ghosh, M. Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* 99–111 (Springer Singapore, 2020).
46. Maxwell, A. E., Warner, T. A. & Fang, F. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* **39**(9), 2784–2817 (2018).
47. Singh, A., Thakur, N. & Sharma, A. A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (INDIACom)* 1310–1315 (Ieee, 2016).
48. Sagi, O. & Rokach, L. Ensemble learning: A survey. *Wiley Interdiscip. Rev.* **8**(4), e1249 (2018).

Author contributions
The authors contributed equally to this work.

Funding
This work was supported by the ENGINEER project. This project has received funding from the European Union's Horizon Europe Framework Programme (HORIZON-WIDERA-2021-ACCESS-03, Twinning Call) under the grant agreement no. 101079377 and the UKRI under project number 10050486. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the UKRI. Neither the European Union nor the UKRI can be held responsible for them.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to E.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.