Machine learning methods for low-cost pollen monitoring - Model optimisation and interpretability

Mills, Sophie A; Maya-Manzano, José M; Tummon, Fiona; MacKenzie, A Rob; Pope, Francis

DOI:
10.1016/j.scitotenv.2023.165853
10.1016/j.scitotenv.2023.165853

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Mills, SA, Maya-Manzano, JM, Tummon, F, MacKenzie, AR & Pope, F 2023, 'Machine learning methods for low-cost pollen monitoring - Model optimisation and interpretability', *Science of the Total Environment*, vol. 903, 165853. https://doi.org/10.1016/j.scitotenv.2023.165853, https://doi.org/10.1016/j.scitotenv.2023.165853

Science of the Total Environment 903 (2023) 165853

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Machine learning methods for low-cost pollen monitoring – Model optimisation and interpretability

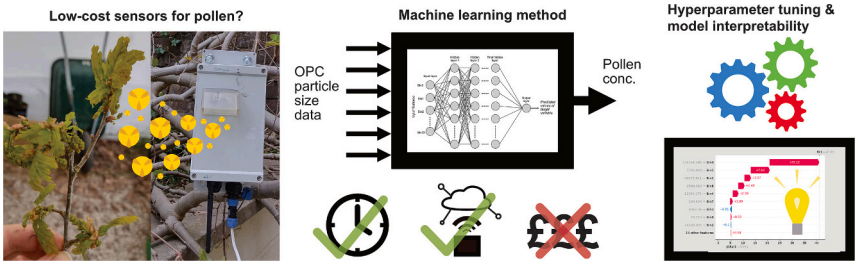
Sophie A. Mills [a,b], José M. Maya-Manzano [c,d], Fiona Tummon [e], A. Rob MacKenzie [a,b], Francis D. Pope [a,b,*]

[a] School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK
[b] Birmingham Institute of Forest Research, University of Birmingham, Birmingham B15 2TT, UK
[c] Centre of Allergy & Environment (ZAUM), Member of the German Centre for Lung Research (DZL), Technical University and Helmholtz Centre Munich, Munich, Germany
[d] Department of Plant Biology, Ecology and Earth Sciences, Area of Botany, University of Extremadura, Badajoz, Spain
[e] Federal Office of Meteorology and Climatology MeteoSwiss, Payerne, Switzerland

HIGHLIGHTS

- Low-cost method for monitoring pollen using machine learning
- Models demonstrated for grass, oak, birch, pine and total pollen
- Methodical neural network hyperparameter tuning for improved model performance
- Use of explainable artificial intelligence analysis to elucidate 'black box' model
- Relationship investigated between different pollen types and observed particle size

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Anastasia Paschalidou

Keywords:
Pollen
Bioaerosols
Automatic monitoring
Low-cost sensors
Machine learning
Explainable artificial intelligence (XAI)

ABSTRACT

Pollen is a major issue globally, causing as much as 40 % of the population to suffer from hay fever and other allergic conditions. Current techniques for monitoring pollen are either laborious and slow, or expensive, thus alternative methods are needed to provide timely and more localised information on airborne pollen concentrations. We have demonstrated previously that low-cost Optical Particle Counter (OPC) sensors can be used to estimate pollen concentrations when machine learning methods are used to process the data and learn the relationships between OPC output data and conventionally measured pollen concentrations.

This study demonstrates how methodical hyperparameter tuning can be employed to significantly improve model performance. We present the results of a range of models based on tuned hyperparameter configurations trained to predict *Poaceae* (Barnhart), *Quercus* (L.), *Betula* (L.), *Pinus* (L.) and total pollen concentrations. The results achieved here are a significant improvement on results we previously reported: the average R2 scores for the total pollen models have at least doubled compared to using previous parameter settings.

Furthermore, we employ the explainable Artificial Intelligence (XAI) technique, SHAP, to interpret the models and understand how each of the input features (i.e. particle sizes) affect the estimated output concentration for each pollen type. In particular, we found that *Quercus* pollen has a strong positive correlation with particles of

* Corresponding author at: School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham B15 2TT, UK.
E-mail address: F.Pope@bham.ac.uk (F.D. Pope).

https://doi.org/10.1016/j.scitotenv.2023.165853
Received 26 May 2023; Received in revised form 10 July 2023; Accepted 26 July 2023
Available online 5 August 2023

optical diameter 1.7–2.3 μm, which distinguishes it from other pollen types such as *Poaceae* and may suggest that type-specific subpollen particles are present in this size range.

There is much further work to be done, especially in training and testing models on data obtained across different environments to evaluate the extent of generalisability. Nevertheless, this work demonstrates the potential this method can offer for low-cost monitoring of pollen and the valuable insight we can gain from what the model has learned.

1. Introduction

Pollen is a major issue globally causing as much as 40 % of industrialised country populations to suffer from hay fever and other allergic conditions (Fröhlich-Nowoisky et al., 2016). Only recently, Japan reportedly 'declared war' on pollen to combat the serious health and quality of life impacts it is having on the population (McCurry, 2023). An important step in addressing this issue is first being able to effectively monitor pollen, and afterwards to map and forecast its spatial distribution as a bioaerosol in the atmosphere. However, the current technology available for doing this is limited and in great need of technological advancements and new innovative methods (Buters et al., 2018; Oteros et al., 2020).

Pollen grains are produced as plant male microgametophytes (containing both vegetative and reproductive cells in varying numbers). Pollen that is most atmospherically relevant, since it is likely to travel the furthest as a free aerosol, would be from anemophilous (wind-pollinated) species. Pollen grains from such species are generally smaller than, for example, insect-pollinated species and have strong association with pollen allergies affecting the public. This includes tree types, such as birch (*Betula*, L.) and oak (*Quercus*, L.), grass (*Poaceae*, Barnhart), and weed types, such as ragweed (*Ambrosia*, L.).

This pollen dispersion accomplishes an important task to ensure successful reproduction when a large portion of it is released into the atmosphere, because they can reach new areas and disseminate their genes. As such, pollen is vital for terrestrial ecosystem regeneration but, during the transport process, it becomes atmospherically relevant too as a class of Primary Biological Aerosol Particle (PBAP). Among PBAPs, which includes fungal spores, bacteria, viruses, and other plant debris (see, e.g., Huffman et al., 2019), pollen grains are generally large in diameter (10–100 μm) and vary widely in shape and size across taxa and species (see, e.g., Després et al., 2012; Reponen, 2011, pp. 723, Bradley, 2015, pp. 408–409).

In the atmosphere, pollen grains can release subpollen particles, for example at high relative humidity conditions or during thunderstorms (Subba et al., 2023). These are often starch granules or other cytoplasmic debris with size ranges reported between 20 nm and 6.5 μm (Stone et al., 2021). Hendrickson et al. (2023) and Matthews et al. (2023) measured wind-stimulated emission factors of subpollen particles from oak (*Quercus*), ryegrass (*Lolium*, L.) and giant ragweed (*Ambrosia*). They reported emissions of 10^3–10^5 particles/pollen grain and 10^{13}–10^{15} particles/m^2 (for a given sample area populated by the species), for particles between 0.010 and 1.0 μm.

Subpollen particles often also carry allergens that affect human health (Bacsi et al., 2006; Smiljanic et al., 2017), while their size also allows them to penetrate more easily into the respiratory system. Meanwhile, pollen and subpollen particles have potential for cloud condensation nucleation (CCN) activity (Pope, 2010; Griffiths et al., 2012; Steiner et al., 2015; Mikhailov et al., 2019) and ice nucleation (IN) activity (Diehl et al., 2001; Diehl et al., 2002; Pummer et al., 2012; Tong et al., 2015; Dreischmeier et al., 2017; Gute and Abbatt, 2020; Burkart et al., 2021) which means they can affect cloud processes, weather and climate.

The current standard methodology for measuring airborne pollen is based on a manual Hirst-type trap to sample particles from the air, then for the particle-impacted melinex tape to be taken to the lab and viewed under the microscope by experienced scientists who count the number of pollen grains present for each type. This method is laborious and time-consuming, and data is not automatically nor quickly available. Other instruments have more recently been developed to automate the process, such as the BAA500 (Oteros et al., 2015), the Rapid-E (Šaulienė et al., 2019), PollenSense (Jiang et al., 2022) and the Swisens Poleno (Chappuis et al., 2020; Huffman et al., 2019) (see Buters et al., 2022, for an overview). Recent work has focused on comparing their performance with the Hirst reference to address the current limitations in pollen monitoring capabilities (Crouzy et al., 2016; Oteros et al., 2020; Maya-Manzano et al., 2023).

Generally, these automated techniques involve machine learning methods, since they are designed to learn from many pollen samples the characteristics of each pollen type through whichever data medium they use. These can include optical light scattering or fluorescence data, but high-end instruments rely largely on image or holographic image data (e.g. Oteros et al., 2020; Sauvageat et al., 2020), i.e. training an algorithm to recognise the individual grain shapes as conventionally experienced scientists do by eye under the microscope for the Hirst. Naturally this requires certain hardware demands and complex computer vision models with stringent sample demands for training. Thus, these instruments can demonstrate good performance with an ability to accurately distinguish pollen grains, however they can also be bulky and expensive.

We recently demonstrated an alternative method for monitoring pollen that is low-cost and convenient in terms of mobility and real-time, remote data accessibility using Alphasense OPC-N3 sensors (see Mills et al., 2023). This method employed supervised machine learning (neural network and random forest) models which were trained, using Hirst data as a benchmark, to produce pollen concentrations from OPC sized particle concentrations. We demonstrated that this method showed promising results and could so far achieve coefficients of determination between model-predicted and Hirst-observed pollen concentrations of up to 0.67, but also highlighted how in particular the neural network methods presented further opportunities for performance improvement.

Machine learning models are optimised to make the best predictions within the feature space representation it is given by the training data. While often demonstrating impressive performance, they can be described as 'black boxes' since it is difficult to truly understand why exactly the predicted outputs are being produced from the input data. This difficulty in human comprehension of the model is because: (a) the model is often too large (i.e., contains too many parameters) to be easily read and interpreted by humans; and (b) the model is not constrained by causal relationships, and so may appear to get the right answer for the 'wrong' reasons (see, e.g., Christianini, 2010; Cristianini, 2023). Being too large to read means that, even though the best description of a model is by definition the model itself, it is frequently useful to have smaller, simpler 'explanation models' that are an interpretable approximation of the original model (Lundberg and Lee, 2017). Having not correctly learned the causal relationships could potentially be an issue when applying a fitted model to a context or environment outside that from which the training data came. A comprehensive review explaining these points in an environmental (ecological) context, can be found by Pichler and Hartig (2023). Thus, for the sake of making decisions based on such models, it appears important to investigate the relationships that the models are learning (we return to this point in the concluding discussion). This can be valuable for deciding how much to trust a certain

model, for investigating potentially unknown variable relationships, or diagnosing models as to which features are serving usefully or perhaps unhelpfully to the model.

There are recently developed techniques for explaining machine learning models and interpreting outputs (often called 'explainable artificial intelligence', or XAI), an overview of which can be found by Gohel et al. (2021). The XAI technique that is utilised in this work is 'SHAP' (SHapley Additive exPlanations), developed by Lundberg and Lee (2017). This is a method inspired by game theory, where Shapley values are calculated to explain the output of a machine learning model by considering all possible combinations ('coalitions') of input features ('players'), quantifying and aggregating the marginal contributions of each feature towards the final output. In this study, we continue to develop the neural network method detailed in our previous work (Mills et al., 2023) for low-cost monitoring of pollen. While the machine learning methods yielded the best results out of those trialled in our study, these methods are not without limitations and should be used only with appropriate consideration of the context. As discussed, these methods provide a trade-off, casting aside interpretability of causal inference in favour of optimal predictive performance based solely on the statistical patterns the model learns (Christianini, 2010). The learning capability of these algorithms, especially deep learning neural networks, depend heavily on data quantity and representative quality. The dataset used in this study is relatively small (6220 data points) for deep learning purposes and covers various pollen seasons that each are absent, emerge, disappear, and overlap for different segments of the time series. Meanwhile, the sensors we use are designed to count general aerosols (i.e. particulate matter) and the task we give the algorithm is to isolate and measure pollen, a small fraction of total ambient aerosols. A further limitation is that the benchmark concentrations we use to teach the algorithm have their own associated uncertainties (Adamov et al., 2021), yet is the best standard we currently have for determining pollen concentrations.

Despite these limitations, our previous work (Mills et al., 2023) showed that machine learning methods are able to learn useful information from this data for public health purposes. In particular, we demonstrated that, on classifying pollen concentrations into 'high' and 'low' categories, we can achieve promising accuracy (F1 scores) with a low false negative rate.

Machine learning has become such a popular and broad field that yields impressive results because the algorithms are generic and can be chosen and applied for a huge variety of tasks, yet come with great flexibility and potential for external optimisation of hyperparameters (on top of the intrinsic optimisation of parameters the model performs itself on training). Control of hyperparameters can determine whether the model underfits, overfits or finds the sweet spot in between. A model underfits when it has low complexity and does not learn effective patterns from training data. It is quite possible for machine learning algorithms with high complexity to overfit, by learning the training data too well that it generalises poorly for unseen (i.e. test) data. Hyperparameters can be used to improve model learning capability but also for regularisation. There are various established regularisation techniques, which can be controlled by hyperparameters and constrain model complexity in a flexible way that adjusts to the input data. Further explanation on the roles of specific hyperparameters relevant to this study can be found in the Supporting Information extended methods section.

In this work, we apply a methodological approach to tune various hyperparameters, including regularisation techniques, with an aim to improve predictive performance and generalisability on unseen data. (In this case, we measure generalisability as being able to make accurate predictions on a hold-out test dataset from the same experimental context. We do not test generalisability of the method on data from a different campaign context.) Selecting the optimal model configuration, we demonstrate the significant improvement in performance metrics this yields. We also extend the method to further pollen types, including *Quercus* (oak), *Betula* (birch) and *Pinus* (L., pine), beyond those reported in the previous work ('total' pollen and *Poaceae* (grass)). The XAI SHAP method is used to interpret and diagnose models for each pollen target variable, to assess which particle size ranges from the OPC are responsible for 'pushing up' or 'pulling down' predicted outputs. We investigate the relationship between OPC bin particle concentrations (input feature values), impact on model (SHAP values) and events of high and low pollen concentration when the model performed with high and low accuracy. We conclude by revisiting the rationale for ML in this domain and outline a pathway for further improvement.

2. Methods

2.1. Context and instrumentation

The context and data used for this study are the same as detailed in previous works, Maya-Manzano et al. (2023) and Mills et al. (2023), from the EUMETNET AutoPollen – ADOPT COST Action (CA18226) Intercomparison Campaign 2021 at the Centre of Allergy & Environment (ZAUM) in Munich, Germany. The data input given to the machine learning models from which to predict pollen concentrations was obtained collectively from three Alphasense OPC-N3 sensors logging data between 9th March and 7th July 2021.

These commercially available OPC sensors count and categorise the size of particles as they pass through a 685 nm-wavelength laser beam, making measurements based on scattered light intensity and Mie scattering theory. A refractive index of 1.5 and particle density of 1.65 g mL^{-1} are assumed. Particle counts are monitored for 24 different bins of varying size ranges between 0.34 and 40 μm. Previously the sensors have been widely used for the monitoring of particular matter air pollution in indoor and outdoor locations (e.g. Crilley et al., 2020; Bousiotis et al., 2023). The sensors were functionalised for outdoor and automatic use, as described in our previous work (Mills et al., 2023), and particle number concentrations (grains m^{-3}) were calculated from raw particle counts for each bin using measured sample period (mean: 5.0 s) and flow rate (mean: 5.2 L min^{-1}) values. The corresponding size ranges for each of the OPC bins can be found in Table S1 in the Supporting Information.

The target data given to the models, to calculate the error of predicted outputs from 'actual' values, was averaged from four Hirst-type samplers (Burkard Manufacturing Co Lid, Rickmansworth, UK), collocated with the OPC sensors. The Hirst data obtained for the intercomparison campaign included number concentrations for 17 pollen types - *Alnus* (Mill., alder), *Ambrosia* (L., ragweed), *Artemisia* (L., mugwort), *Betula* (L., birch), *Carpinus* (L., hornbeam), *Corylus* (L., hazel), *Fagus* (L., beech), *Fraxinus* (Tourn. ex L., ash), *Picea* (A. Dietr., spruce), *Pinus* (L., pine), *Plantago* (L., plantain), *Poaceae* (Barnhart, grass), *Populus* (L., poplar), *Quercus* (L., oak), *Taxaceae-Cupressaceae* (Gray, yew), *Tilia* (L., lime), *Urtica* (L., nettle) – and varia (other pollen that were not included in the previous list or unknown pollen) and for 'total' pollen which was the sum of all types present at any given time. Full details on the context of the whole campaign can be found in Maya-Manzano et al. (2023) and further details on the Hirst observations are available in Triviño et al. (2023).

2.2. Neural network hyperparameter tuning

Scripts for external (hyperparameter) optimisation were run in Python 3.9.6 using the University of Birmingham's BlueBEAR Supercomputing facility. The module versions Scikit-learn 1.0.1, Tensorflow 2.8.4, Matplotlib 3.4.3, Numpy 1.21.3 and Pandas 1.3.4 were utilised. Model architecture was initially constructed with the same hyperparameters as demonstrated in our previous work (Mills et al., 2023) and then a range of values for selected hyperparameters were tested. However, this time the original dataset was split into training, validation and test datasets (as opposed to just training and test datasets). The

training dataset was used for the model to directly train on. The validation dataset was used to determine a validation score by the loss metric (Mean Squared Error (MSE)) for a hold-out dataset for each training iteration (epoch) alongside a training score. The validation score was used to inform the model when to stop updating the parameters, i.e. when the mean squared error (between predictions and observations) had stopped decreasing further. The test dataset was used to calculate metrics, including Root Mean Squared Error (RMSE) and R^2 score (coefficient of determination), to evaluate the various models and effect of varying hyperparameters. The external optimisation procedure was computationally expensive, involving the training of many models, and therefore was only applied to one input feature-target variable pairing as a demonstration for this work. Thus, the specific optimisation results reported here apply to the model which is trained to predict total pollen concentrations from the 24 particle size bins available from the OPC data (i.e. not including temperature and relative humidity). We assume going forward that the model tasks are generally similar enough to benefit from hyperparameters tuned for just one of the models.

The tested hyperparameters were as follows: number of layers and number of nodes in each layer (i.e. model capacity), batch size, optimisation algorithm, initial learning rate of best optimisation algorithm, activation function at hidden layer nodes, scaling of target variable, train:validation:test datasets split size, proportion of dropout nodes in each layer, weight constraints applied in each layer, and implementation of noise at different configurations within the network. A summary of the values trialled for each hyperparameter is presented in the table in Table 1.

An extended methods section containing further details on each hyperparameter and the configurations chosen for this study is provided in the Supporting Information. Further explanations on the effect each hyperparameter has on a given model and other general information can be found in the book 'Deep Learning' by Goodfellow et al. (2017) and ebook 'Better Deep Learning' by Brownlee (2018) as well as other resources by Bengio (2012a, 2012b), Masters and Luschi (2018), and Reed and Marks (1999).

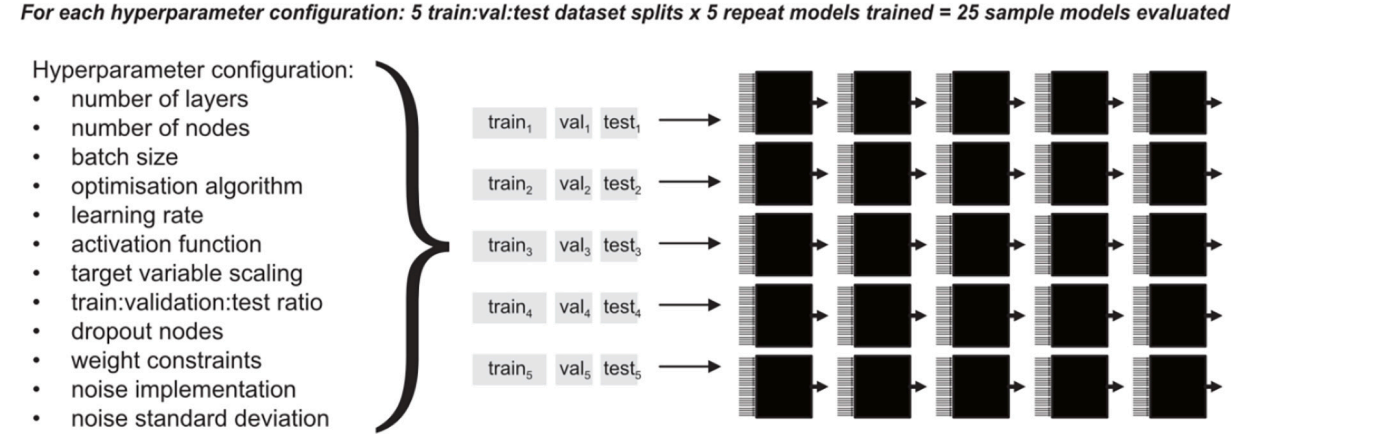
For testing, 5 distinct splits of training, validation and test data were created randomly (with the same random seeds for each hyperparameter test) from the whole dataset. Models were then trained and evaluated 5 times on each of the 5 sets of train-validation-test data for each hyperparameter configuration. Thus, each hyperparameter configuration was sampled 25 different times in total, over 5 different combinations of training, validation and test datasets (visualised in Table 1). Due to the stochastic nature of machine learning algorithms they can produce different results on multiple runs, and they are also sensitive to the provided train/test data. We followed this method to allow for variation caused by differences in train/test datasets and the stochastic process to truly assess the effect of varying each hyperparameter.

A metrics dataset was produced from each hyperparameter experiment – including RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), Spearman correlation coefficient values and R^2 scores for model-predicted values vs Hirst-measured pollen concentrations based on the test dataset in each case. These results were analysed and visualised using Python 3.9.7 in Jupyter Notebook from the Anaconda Distribution with module versions Numpy 1.24.1, Pandas 1.5.3, Matplotlib 3.6.3 and Seaborn 0.12.2. Box and whisker plots were produced to summarise the R^2 scores achieved across all 25 sample models for each hyperparameter configuration for each experiment. R^2 scores were chosen as the primary metric for comparison, since they provide better

Table 1

Above: Summary table of hyperparameter configurations trialled for this study. *Initial learning rate was only trialled for chosen optimisation algorithm (Adam). Below: Figure showing the method by which the various hyperparameter configurations were tested. Five repeat models were trained for five different train:validation:test dataset splits, resulting in a total of 25 sample models for each hyperparameter configuration.

Hyperparameter	Trialled values
Number of layers	3; 4; 5; 6; 7
Number of nodes	50; 100; 200; 400
Batch size	32; 64; 128; 256; 512
Optimisation algorithm	*RMSProp* (Root Mean Square Propagation); *Adagrad* (Adaptive Gradient Algorithm); *Adam* (Adaptive Moment Estimation);
Initial learning rate*	0.0001; 0.0005; 0.001; 0.005
Activation function	ReLU; ELU; SELU
Scaling of target variable	True; False
Train:Validation:Test dataset ratio	50:25:25; 60:20:20; 70:15:15; 80:10:10
Dropout node proportion	0.1; 0.2; 0.3; 0.4; 0.5
Weight constraint	Max-norms 1.0; 2.0; 3.0; 4.0; 5.0 and unit norm.
Noise implementation	None; input layer; hidden layers
Noise standard deviation	0.001; 0.0001; 0.00001



comparison across models with different target value ranges than RMSE or. This was also the metric used to assess the performance of all instruments that took part in the AutoPollen intercomparison campaign (see Maya-Manzano et al., 2023).

2.3. Final models

The results from the hyperparameter tuning experiments were used to inform the configurations applied to the final optimised models. Final models were trained for 3 sets of input features – OPC bins only, OPC bins, RH and temperature, and RH and temperature only – and 5 different target variables – total pollen, *Poaceae*, *Quercus*, *Betula* and *Pinus* pollen categories. These pollen types were chosen for their relevance to public health (see, e.g., Darrow et al., 2012) and their prevalence in the dataset of this environment. The same hyperparameter configurations were used for all models, with the exception of train: validation:test ratio, which was 80:10:10 for total pollen and *Poaceae* models and 60:20:20 for the rest. This was because, although the 80:10:10 split yielded optimal performance for the total pollen model that underwent hyperparameter tuning, for *Quercus* and other less prevalent types in this case the test dataset seemed not to be large enough to be representative. A greater test dataset size reduces the proportion that can be used for training but ensures that the model is evaluated fairly on a representative sample.

This combination of input features and target variables produced 15 different model sets, each of which was trained for 5 sample models within each set. The model hyperparameters within each set were identical however, due to the stochastic nature of the learning process, the 'optimal' model weights (parameters) settled on in each case will be different depending on the statistical patterns each model has learned. R^2 scores were calculated for all models between relevant target variables and model predictions from the unseen test dataset. The resulting spread of R^2 scores for each model set was visualised and the mean and best scores recorded. We can say that models which achieved high R^2 scores generalised well in terms of the hold-out test dataset here, but we cannot comment on how well they would generalise in other contexts. For example, the models with the best R^2 scores here may not be those with the best R^2 (and therefore generalisability) in another context, though this is unlikely to be models with very low scores here. If the spread of scores among a model set is narrow, it suggests any model with these hyperparameter configurations will result in similar performance and hence these hyperparameter configurations may be more robust. If there is great variance among R^2 scores, it may suggest the hyperparameter configuration is not very robust and learning outcome depends heavily on the stochastic learning process.

This process was also applied to produce an equivalent set of models, predictions and R^2 scores for models trained on daily (24 h) time resolution data. This was to observe the difference in performance between models trained on hourly and daily resolution data. It was also in preparation to apply for future comparison purposes, since pollen concentration data from the standard Hirst-type samplers is often only available in daily time resolutions.

2.4. Model interpretation using SHAP values

To explain the models, understand the input feature-output variable relationships and why certain predictions are produced, we employed the SHAP (Shapley Additive exPlanations) method developed by Lundberg and Lee (2017). SHAP uses Shapley values from game theory to explain the output of any machine learning model. It does this by considering all possible combinations, or 'coalitions', of input features ('players' from a game theory perspective), quantifying the marginal contributions each feature makes towards the final output when added to coalitions, then aggregating the marginal contributions. The sum of SHAP values across all features for any given observation equates to the difference between the model prediction and the observation mean, or 'null model'. This process involves training models for each distinct feature coalition to calculate differences in predicted output (marginal contribution) and so is computationally expensive. SHAP values have limitations when feature dependencies are present and cannot simply be taken for causal inference; it tells us how important the feature is to the model. Nevertheless, it is a powerful method for interpreting machine learning models and can be implemented efficiently using Lundberg's SHAP library (https://github.com/slungberg/shap).

We took the best performing models (highest R^2 score on test dataset) for each of our model sets and applied the SHAP method to investigate how each of the given features were affecting the output of the model. 'Bee swarm' plots were constructed in each case, which give information on feature importance, magnitude of the affect each feature has on the model output, and the corresponding magnitude of the input feature value in each case. The information from these plots was condensed into a matrix plot which scored input feature and target variable pairings by colour for strong positive correlation, weak positive correlation, strong negative correlation, weak negative correlation and no correlation respectively. Correlations were nonlinear and decided for each case based on the visual distribution of data points on the plot demonstrating feature value and SHAP value (i.e. impact on output prediction of target variable).

To investigate further, we isolated data points that fit into four categories: high pollen low error (HPLE), low pollen low error (LPLE), high pollen high error (HPHE) and low pollen high error (LPHE) events - after excluding all zero values for the target variable. High and low pollen was defined as all data points above the 75th quantile and below the 25th quantile for the Hirst-measured target pollen variable and high and low error was defined as above the 75th and below the 25th quantile for the absolute difference between Hirst-observed and model-predicted target variable concentrations. The only exception for this was that for *Betula* the low error threshold was raised to the 40th quantile for the HPLE category because there were no data points available satisfying the 25th quantile threshold for high ($>$ 75th quantile) *Quercus* concentrations. The percentage of non-zero values for each target variable in the whole and test datasets, the number of data points that were filtered by each category threshold, and the specific quantile threshold values calculated in each case can all be found in the SI in Tables S3-S5.

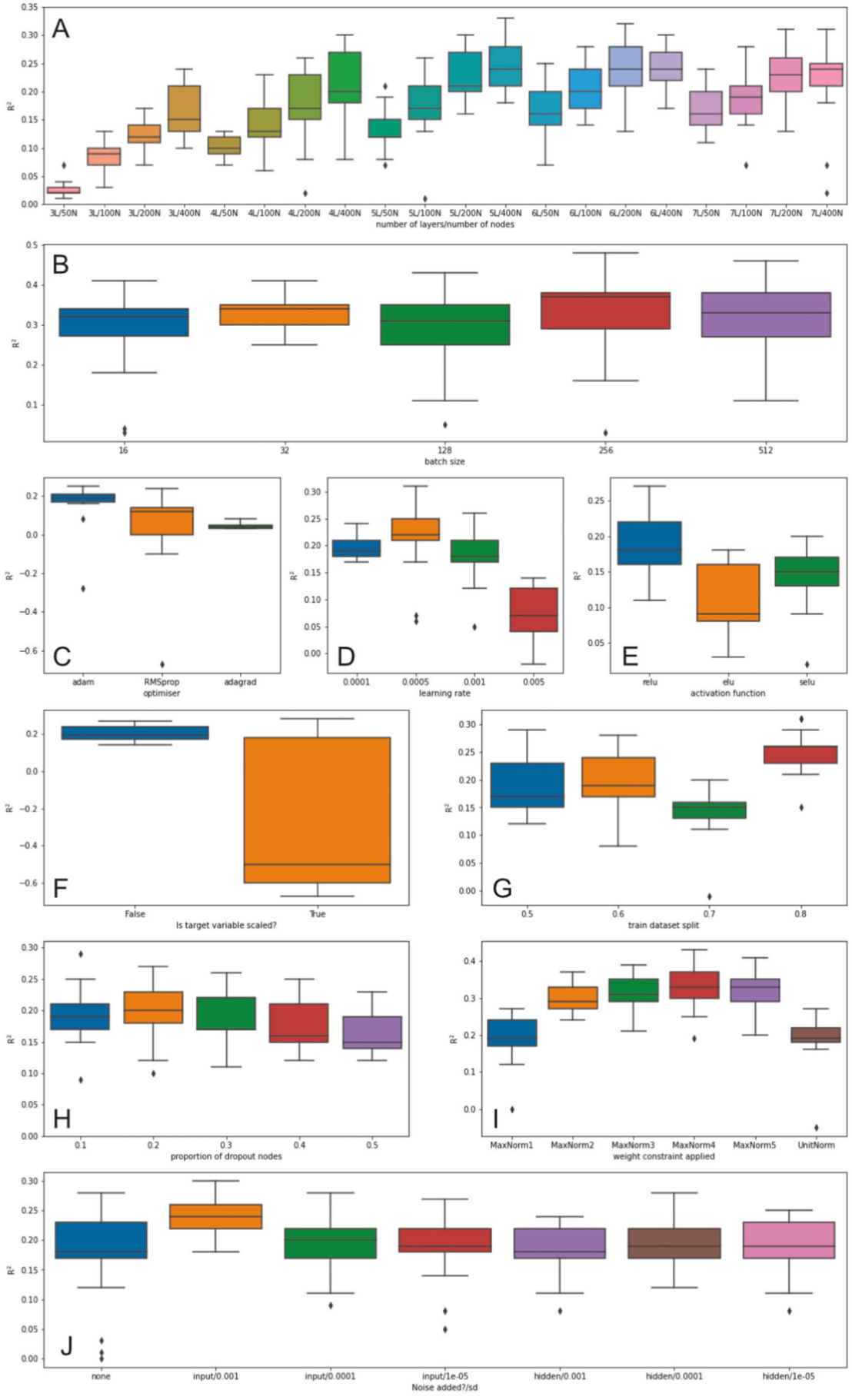
The original feature (bin) values and SHAP values were averaged over all data points present in each group. Bar plots were constructed to show the relative averaged feature values for each category, and a colour scale added corresponding to SHAP value sign and magnitude.

3. Results and discussion

3.1. Hyperparameter tuning

The results of the optimisation experiments are summarised in Fig. 1, based on the R^2 scores calculated from the test datasets across the 25 samples for each configuration. This visualises the effects of varying the model capacity, by the number of layers and nodes per layer in the neural network architectures, and model learning behaviour by varying batch size, optimisation algorithm, learning rate and node activation functions. Also visualised are the effects of certain data preparation considerations, such as scaling of the target variable and the test-train split size, as well as the effects of varying regularisation methods to improve generalisation such as dropout percentage, weight constraints and addition of noise.

In theory, there would be an optimal capacity of the model architecture, defined by the number of layers and nodes, with enough complexity to address the given task yet not so much as to learn the training dataset too meticulously (i.e., to be fitted but not overfitted). Increasing the number of layers is computationally cheaper than increasing the number of nodes in a layer, yet the optimal capacity in practice can only be found empirically, keeping computational limitations also in mind. The results from our empirical experiments here in



(caption on next page)

Fig. 1. Box and whisker plots showing effect of varying various hyperparameters on model performance, quantified by the R^2 score (coefficient of determination). Plot A: model capacity, L = number of layers and N = number of nodes. Plot B: batch size. Plot C: optimisation algorithm. Plot D: initial learning rate for Adam optimisation algorithm. Plot E: activation function applied at hidden layers. Plot F: target variable scaling. Plot G: train:validation:test split ratio – x-axis value corresponds to the fraction of the whole dataset used for training, while validation and test datasets where taken in equal parts from the remaining fraction. Plot H: dropout node proportion. Plot I: weight constraint. Plot J: noise implementation (none/input layer/hidden layers) and standard deviation applied. Colours are unique for each x-axis/hyperparameter value.

plot A of Fig. 1 demonstrate the general increase in model performance with increasing number of layers and nodes, but the increase with additional layers diminishes after 5 layers for a large number of nodes ($N \geq 200$). The configuration that achieved the highest mean R^2 score (0.24) across all model samples was the combination of 5 layers and 400 nodes; thus this combination was chosen to proceed. (Other combination means also achieved close to this (>0.22), e.g. 6 or 7 layers with 200 or 400 nodes, but these were also more computationally expensive.)

Batch size dictates how often the weights of the model are updated during training, thus affecting training dynamics. No clear trend was observed across different batch sizes. The highest mean across models (see plot B in Fig. 1) was achieved with a batch size of 256 (0.22), however, the difference among other batch size means was not large (<0.11). The results using the commonly used batch size of 32 showed the smallest range and highest minimum, and therefore was chosen to proceed for further models.

The *Adam* optimiser ($R^2 = 0.17$) (see plot C in Fig. 1) clearly performed better than the *RMSprop* ($R^2 = 0.06$) or *Adagrad* ($R^2 = 0.04$) alternatives, when used on default settings. When the initial learning rate was varied for the *Adam* optimiser (see plot D in Fig. 1) an optimal value of 0.0005 was found ($R^2 = 0.22$). The standard, and less computationally expensive, ReLU activation function applied at each hidden layer in the network proved to yield generally higher R^2 scores (0.19) than either of ELU (0.11) or SELU (0.15) (see plot E in Fig. 1). This suggests that the extra nuance these activation functions can add (e.g. decreasing bias shift, self-normalisation, and robustness – see Pedamonti, 2018; Marchisio et al., 2018) did not offer an advantage. On the contrary, it appears that the characteristic of the ReLU activation function to equate all negative incoming values to zero (while ELU and SELU continue to propagate negative values) has benefited the algorithm in this case. The best performing configuration in each case was taken to proceed.

Previous work during the development of these models (in Mills et al., 2023) showed that scaling (normalisation) of the particle concentration input data is important for successful model training with good predictive performance. However, it was not previously conclusive whether scaling (normalising) the pollen concentration target variable was useful or not. The results here (see plot F in Fig. 1) demonstrate that scaling the target variable in this case is not useful ($R^2 = -0.24$), but rather leaving it unscaled results in generally better (0.20) and less variable performance.

While some testing had been tried before, we decided to investigate more thoroughly the potential effect that the train-test split size had on estimated model performance. We considered this useful to test as the available data here is limited and therefore wanted to make best use of the available data to provide effectively for both model training and evaluation. The varied proportions of the whole dataset taken for the training dataset are labelled in the x-axis of plot G in Fig. 1. It is important to note that the remaining proportion of the dataset was then split once more, in equal parts, into validation (used to settle on the optimal model over training epochs) and test datasets. The latter of which was used for final model evaluation.

From the results, we found that using 80 % for the training dataset, and therefore 10 % each for validation and test datasets, yielded the best performance ($R^2 = 0.25$). This split was used for further models predicting total pollen and *Poaceae* pollen, however it was observed that this did not generally result in best performance for models targeting *Quercus* pollen. This is likely because the representation (i.e. season length) of the *Quercus* pollen season in the given dataset is relatively small, at least compared to 'total' or *Poaceae* pollen, so a smaller dataset proportion used for evaluation would less likely be representative. After testing splits of 80:10:10 and 60:20:20 (train:validation:test) it was decided that the latter was best when applied to *Quercus*, *Betula* and *Pinus* models. (This split ratio also achieved the second best R^2 score (0.20) for the original total pollen hyperparameter testing experiment.)

Dropout layers and weight constraints are both effective methods for model regularisation and reducing overfitting, but again the optimal tuning of dropout node percentage and threshold values for weight constraints can generally only be determined empirically. We found for the purposes of our task that a dropout percentage of 20 %, see plot H in Fig. 1, and weights constrained by a max-norm (upper value threshold) of 4.0, see plot I in Fig. 1, at each hidden layer yielded optimal performance (R^2 scores of 0.20 and 0.33 respectively). These configurations were applied to further models.

Finally, the addition of a small amount of noise at input or hidden layers was trialled to see if it affected the ability of the model to accurately predict the unseen test dataset (in theory by limiting dependency on training data and improving generalisability). Our results found that the addition of noise to the data at the input layer with a standard deviation of 0.001 improved model performance on the test dataset ($R^2 = 0.24$), compared to no noise being added ($R^2 = 0.19$), and among all other tested noise configurations. Thus, this addition of noise to input data was also applied to further models.

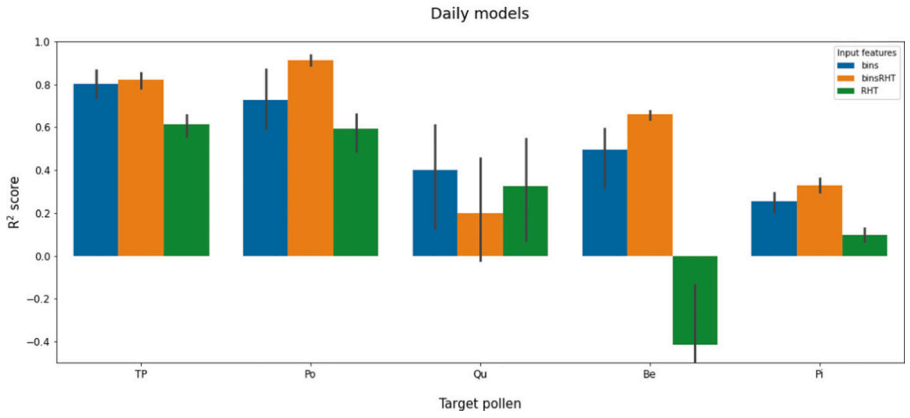
3.2. Improved models from chosen hyperparameter values

The plots in Fig. 2 present the results of the final model evaluations after the optimisation process for hourly and daily resolution models respectively, while the mean and best R^2 scores achieved for each model set are summarised in the table in Fig. 2. We evaluate, in particular, the models that used OPC particle size bins data since it is that sensor-ML combination that we are primarily concerned with. The models based solely on meteorological variable input feature are provided for comparison to demonstrate the importance of the particle size bin data for the models.

The results achieved here are a significant improvement to the initial results reported by Mills et al. (2023), denoted by the horizontal white dotted lines. The average R^2 scores for the total pollen models have at least doubled - from 0.15 to 0.49 without meteorological variables and 0.3 to 0.65 with them - and *Poaceae* models improved from 0.34 to 0.48 without, and 0.67 to 0.87 with, meteorological input features. The greater improvement in total pollen models demonstrates the gain of hyperparameter tuning for specific tasks. These results also demonstrate the method being applied to a greater variety of pollen types, which appears to show similar promise provided appropriate data availability and model optimisation.

From the hourly and daily model sets (blue and orange only), 6/10 and 8/10 respectively produced a model that achieved an R^2 score ≥ 0.5 and there are even some models that have performed extremely well ($R^2 \geq 0.8$) when evaluated on the test dataset. The daily resolution models generally outperformed the hourly resolution models. It has been reported previously (Maya-Manzano et al., 2023) that comparison alongside the Hirst-type sampler generally yields better results at coarser time resolutions, since the temporal averaging results in less deviation between instruments.

Models targeting total and *Poaceae* pollen generally performed better



Model set	Hourly mean R^2 score	Hourly best R^2 score	Daily mean R^2 score	Daily best R^2 score
binsTP	0.49	0.50	0.80	0.90
binsRHTTP	0.65	0.73	0.82	0.87
RHTTP	0.36	0.39	0.61	0.68
binsPo	0.48	0.53	0.73	0.95
binsRHTPo	0.87	0.89	0.91	0.95
RHTPo	0.45	0.50	0.59	0.67
binsQu	0.31	0.36	0.40	0.67
binsRHTQu	0.45	0.62	0.20	0.66
RHTQu	0.09	0.22	0.33	0.66
binsBe	0.31	0.36	0.50	0.61
binsRHTBe	-0.53	-0.05	0.66	0.68
RHTBe	-0.05	0.04	-0.42	-0.01
binsPi	0.36	0.38	0.25	0.31
binsRHTPi	0.68	0.73	0.33	0.38
RHTPi	0.06	0.16	0.10	0.15

Fig. 2. Above: Summary of R^2 scores achieved for final hourly resolution models. Horizontal white dotted lines denote scores achieved in previous study for models tested there (Mills et al., 2023). Middle: Summary of R^2 scores achieved for final daily resolution models. Below: Summary of R^2 scores achieved for all final models, both hourly and daily time resolution. TP = total pollen; Po = Poaceae; Qu = Quercus; Be = Betula; Pi = Pinus. Blue = models with only OPC bin input features; orange = models with OPC bin & meteorological input features; green = models with only meteorological input features.

than others. This is likely due to the better representation of these target variables within the time period for which data was collected and the fact that a larger proportion of the dataset was used for training (80 % as opposed to 60 % for the rest). In general, as more data is fed to machine learning models to learn from their prediction performance increases. However, we had a limited supply of data here. While total pollen models demonstrate the most improvement, since this target variable was used for hyperparameter tuning, it is interesting that the task of predicting Poaceae pollen also benefited from this process. The other types may also have benefitted to an extent, yet these models could benefit further by tuning hyperparameters for each pollen type.

There were some instances, in particular for the *Betula* model with meteorological input features, where the R^2 score was small or negative, implying that these models did not learn reliably useful information. This may be possible to improve upon by applying hyperparameter tuning for this specific model case, or it may simply be due to limitations of the available data. However, since the equivalent daily-resolution trained model appears to have performed considerably well, it does suggest there is useful information from which to learn the *Betula* type.

In the overview study which summarised the blind performance of all instruments participating in the intercomparison campaign (Maya-Manzano et al., 2023), an R^2 of >0.5 was considered good while >0.75 was considered excellent agreement with the benchmark Hirst instruments. Measuring total pollen, 9/18 of the automated pollen monitoring systems achieved $R^2 > 0.5$ for 3-hourly and daily resolution, while 3/18 achieved $R^2 > 0.75$ for daily resolution. Those instruments that performed best included the BAA500 and the Swisens Poleno. These instruments (depending on the algorithm applied) also demonstrated R^2 scores above 0.5 for daily and above 0.75 for 3-hourly *Poaceae* pollen concentrations. R^2 scores of above 0.75 were observed for 3-hourly and daily measurements of *Quercus* and *Betula* pollen, the latter with the most success.

Our results from this study applying low-cost sensors seem impressive, with over half of our models achieving $R^2 > 0.5$ and some even >0.75. However, it should be noted that these algorithms are not blind and have been specifically trained on the data in this context. Meanwhile, the results reported in the overview paper are for instruments that use machine learning algorithms trained independently from the Hirst data, on specific pollen samples previously provided to them.

3.3. Model interpretation and input-output relationships

Fig. 3 shows bee swarm plots from the Python SHAP package for the test datasets for each pollen type. The relationships learned by the model are generally not linear and are complex to interpret (see Introduction, above, and Lundberg and Lee, 2017), but our aim here is to discern which particle size ranges generally increased when a given pollen type was present (i.e., find and interpret an 'explanation model' for that purpose). Strong SHAP signals may relate to the size ranges of subpollen particles but we have no way of demonstrating this. We also aim to distinguish differences among the types that the model has learnt, in order to assess the feasibility of differentiating between types. The general observed correlations based on the bee swarm plots are summarised in the matrix plot in Fig. 3. Corresponding bee swarm plots for models based on OPC bins and meteorological variables are displayed in Fig. S1 and partial dependence plots for each input bin and target pollen type can be found in Fig. S2–6 in the SI.

While the meteorological variables were generally very useful for predicting pollen concentration, for models where they were included the correlations among the OPC bin ranges appear less clear. Therefore, for investigating the impact of the different bin sizes in each case we have focused on the models trained on OPC bins only.

Fig. 4 shows bar plots visualising relative averaged particle concentrations for each OPC bin for four categories of high and low pollen concentrations and error. Bins beyond bin 11 (> 10 μm) have not been included since few particles were detected in that size range and the bee swarm plots did not provide useful information. Generally, when particles were observed in these larger-size bins, it pushed the predicted output of the model down, as can be seen for larger-size bins for the total pollen, *Poaceae*, and *Betula* plots.

These plots can give further evidence about which features are affecting the model and are correlated with certain events such as high pollen. For example, a taller bar of a strong red colour for the HPLE category (e.g. bin 0 for *Poaceae*) compared beside a shorter bar of a blue colour for LPLE suggests that high values of this feature contributed greatly to low error high pollen events while comparatively smaller values were also indicative of low error low pollen events. We may also be able to suspect features that might be confounding the model and be responsible for inaccurate predictions. In particular, strong blue bars for the HPHE category may suggest this feature is pulling the model predictions down and creating a large error from the high output value it should be. Likewise, strong red bars for the LPHE category (e.g. bin 5 for *Pinus*) may suggest the feature is pushing the model predictions up even when the actual pollen concentrations are low, potentially confusing the model.

3.4. Total pollen

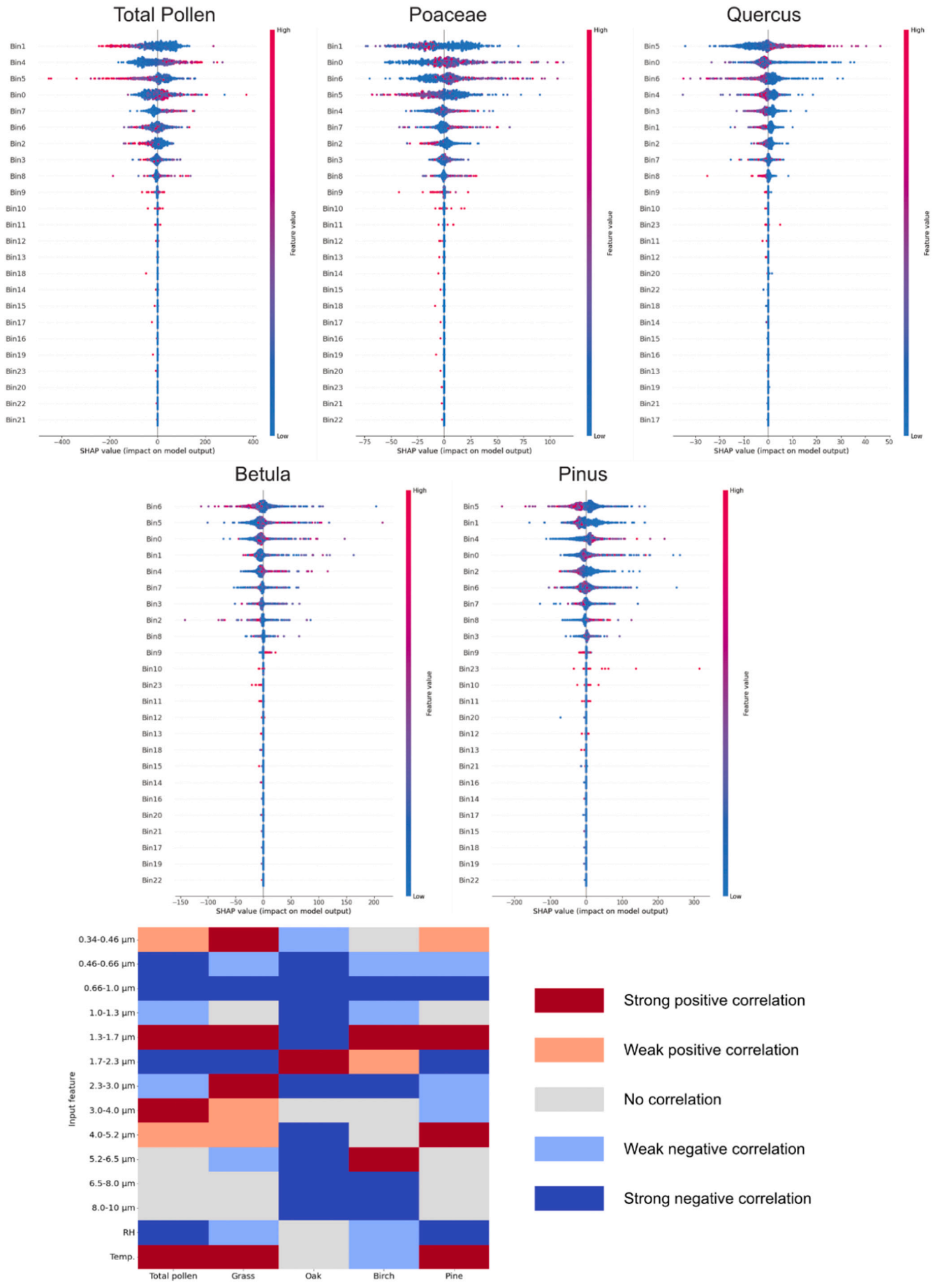
From the bee swarm plots in Fig. 3, it appears that total pollen had some positive correlation with bins 0, 4, 7 and 8 (corresponding particle sizes can be found in Table S1 in the SI). From the bar plots in Fig. 4, comparing high/low pollen events with high/low error it appears that bins 0–3, 5 and 9–11 had large responsibility for pushing the predicted pollen concentration down when particles in these bins were at high concentrations. Meanwhile bins 6–8 are responsible for pushing the pollen concentrations up with high particle concentrations present in these bins. We conclude that total pollen here is largely correlated with particles, measured optically by the OPC, in the size range 2.3–5.2 μm.

The model appears to have associated bins 1 and 5 with a strong negative correlation to total pollen. For example, when these particle concentrations are high they have contributed strongly to predicting low pollen events (i.e. strong blue bars for LPLE). However, the issue is that when the pollen concentrations have been low it has contributed strongly to predicting that pollen is high when it is actually not (i.e. strong red bars for LPHE). This suggests that these bins are causing the model to learn unhelpful correlations and it might be beneficial to remove them as input features.

The total pollen target variable encompasses a wide range of pollen types which may be associated with varying particle sizes and would be very dependent on the representation of 'total pollen' in this context (e.g. majority of *Poaceae* representation in this context). Therefore it can be expected that there may be conflicting correlations with different types within this target variable and the model may not generalise well in other contexts where the representation of 'total pollen' is different. It could be expected to generalise less well than other specific pollen type models in other contexts.

3.5. Poaceae

From the bee swarm plots, the *Poaceae* type appears to have positive correlation with bins 0, 4, and 6–8. From the bar plots, bins 0 and 6–8 are particularly congruent with this positive correlation. This suggests that particles in size ranges 0.35–0.46 and 2.3–5.2 μm are largely co-existent with *Poaceae* pollen type in this context, similar to the observed size range for total pollen. Bins 1 and 5 also show a similar effect as described for total pollen, being responsible for the high error low pollen (LPHE) events. These similarities are due to the fact that the total pollen model in this context is largely represented by the *Poaceae* type.



(caption on next page)

Fig. 3. Above: SHAP 'bee swarm' plots for all pollen types bins-only models. The features are ranked in order of general importance and each point corresponds to a datapoint in the test dataset. The x axis corresponds to the impact each datapoint and feature had on the model output (i.e. target pollen concentration), relative to the mean output at 0. The colour corresponds to the actual value of that feature – red for high particle concentrations in a given bin and blue for low values. Therefore, feature rows which generally have more red further to the right and blue further to the left on the x axis likely have some general positive correlation between feature value and SHAP values (impact on the model) and vice versa for negative correlation. Corresponding bee swarm plots for models based on OPC bins and meteorological variables are displayed in Fig. S1 and partial dependence plots for each input bin and target pollen type can be found in Fig. S2–6 in the SI. Below: Matrix plot summarising observable nonlinear correlations between input features and target pollen variables.

3.6. *Quercus*

The *Quercus* model specifically shows strong correlation with bin 5 in the bee swarm plot, while most other bins suggest negative correlation. While it appears many bins have contributed to pushing the *Quercus* model output up generally, the bar plots (Fig. 4) show that specifically bins 4–6 demonstrate a stronger impact on the model (i.e. darker red colour) with higher particle concentrations in these bins (i.e. taller bars) for high pollen low error events (HPLE; 1st bar) when compared to low pollen low error events (LPLE; 2nd bar). This implies that the *Quercus* type is particularly associated with particles in the size range of 1.3–3.0 μm.

The bar plots suggest that bin 0 is responsible for pushing the predicted output concentration up when actual pollen concentrations are low (strong red LPHE bar) and is responsible for the high error in the HPHE category (blue bar, i.e. bringing predictions down when actual pollen is high). Other bins including bins 1–4 and 6 seem also responsible for causing high error for LPHE events. Bin 5 is an example that the model has learned the relationship well, since the bars are red for high pollen events (HPLE and HPHE) and blue for low pollen events (LPLE and LPHE) but the overall output may be confounded by other less helpful bins. Here, for example, it may be beneficial to remove bins 0–3 as input features and allow the model to learn to make better use of bins 4–6.

3.7. *Betula*

The bee swarm plots suggest that the *Betula* type has positive correlations with bins 4,5 and 9. Meanwhile, the bar plots suggest that accurate high pollen events are largely positively correlated with bins 4–7 (i.e. darker red colour and taller bars for HPLE particularly when compared to LPLE), but not bin 9. This suggests that *Betula* can be associated with particles in size ranges 1.3–2.3 μm (bins 4–5), possibly extending up to 4 μm (bins 6–7). This size range is similar to *Quercus* and the two types have similar pollen size ranges - 17-26 μm for *Betula* (Mäkelä, 2009) and 20–42 μm for *Quercus* (Wrońska-Pilarek et al., 2016) – while generally smaller than *Poaceae* (22–46 μm, Radaeski et al., 2016) or *Pinus* (28–97 μm, Song et al., 2012). This may suggest that not only are the intact pollen grains of similar sizes but so are the associated subpollen particles that come from each.

3.8. *Pinus*

Pinus appears to have positive correlations with bins 0, 4 and 7 from the bee swarm plots. This positive correlation is only reinforced for bin 0 in the bar plots, as this is the only case where the bin particle concentration is higher for HPLE events compared to LPLE events while the SHAP values (impact on the predicted *Pinus* concentration) were strongly positive. This corresponds with particle sizes between 0.35 and 0.46 μm however, being the lowest size range measurable by the OPC instrument there may be subpollen particles associated with *Pinus* below this size range. There appear to be some similarities in positively correlated bins when compared with *Poaceae* according to the bee swarm plots which are not reinforced by the high pollen event bar plots. This may be due to the fact that the *Pinus* season overlapped with *Poaceae*, so the two types may often have co-occurred leading to learned associations related to the wrong type, i.e. *Pinus* learning from *Poaceae* subpollen particles.

The most unhelpful feature for the *Pinus* model appears to be bin 5 since it has pulled the predicted output down when it should be high (i.e. blue bar for HPLE) and strongly pushed the output up when it should be low (i.e. strong red bar for LPHE). Meanwhile, even bins that have been helpful for distinguishing HPLE and LPHE events (e.g. bins 0–2) have also contributed to the high error when pollen should be low (red for LPHE). This model may benefit from removing certain features such as bin 5 but also by training on data in a different context that is not confounded by co-occurring pollen types. This would in fact be desirable for each taxon, however, may be hard to realise in practice.

3.9. *Summary*

The conclusions of the results discussed in Sections 3.3–3.8, taking into account the SHAP bee swarm plots in Fig. 3 and the SHAP bar plots in Fig. 4, are summarised visually in Fig. 5. It shows which bin size ranges contributed positively to the detection of each pollen type, and demonstrates the similarities and differences in terms of the associated bins. For example, total and *Poaceae* pollen are similar, due to total pollen being largely represented by *Poaceae* in this dataset, as well as *Quercus* and *Betula* pollen, likely because the pollen grains or associated particles are similar in size.

3.10. *Importance of input features and context*

We proceeded to test the effect of removing bins that may have confounded the model in each case according to the discussed observations. The resulting R^2 scores for each model set with certain bins removed are displayed in Fig. 7 in the SI. Removing those above bin 8 (>5.2 μm) did not significantly affect performance. However, removing other specific bins considered potentially 'unhelpful' in each case (detailed in SI Fig. 7 caption) resulted in decreased performance, generally the more bins removed the worse the score. The models do not seem to gain much information from bins above 5.2 μm, which may be in part due to limitations of the OPC instrument when measuring larger particles. However, the results suggest that the models make use of all other bins for optimal performance in this context, as they are able to learn even very complex nonlinear relationships that are difficult for us to comprehend or visualise.

This may include learned statistical relationships with some size ranges that are collinear with the given pollen target only in this context, and which may become irrelevant and unhelpful if applied to other contexts. If the bins that appear positively correlated with a given target pollen taxon are directly associated with subpollen particles within those size ranges, then these bins may generalise well in other environments where the same taxon is present. However, bins that have simply been used by the model because they have some collinearity with the given target taxon (in particular, this may be relevant for negatively correlated bins) may cause the model to generalise less well in other environments and hence their removal may bring higher scores when applied to different contexts.

Ultimately, it must be emphasised how sensitive such models are to the environment and context from which training data is taken. A model that performs best on the reserved test dataset here may not necessarily perform best on a test dataset in another context. In other words, a model with a lower R^2 score here could potentially generalise better in other environments. Thus, we emphasise the need to further validate these models with more data in different contexts. For such methods to

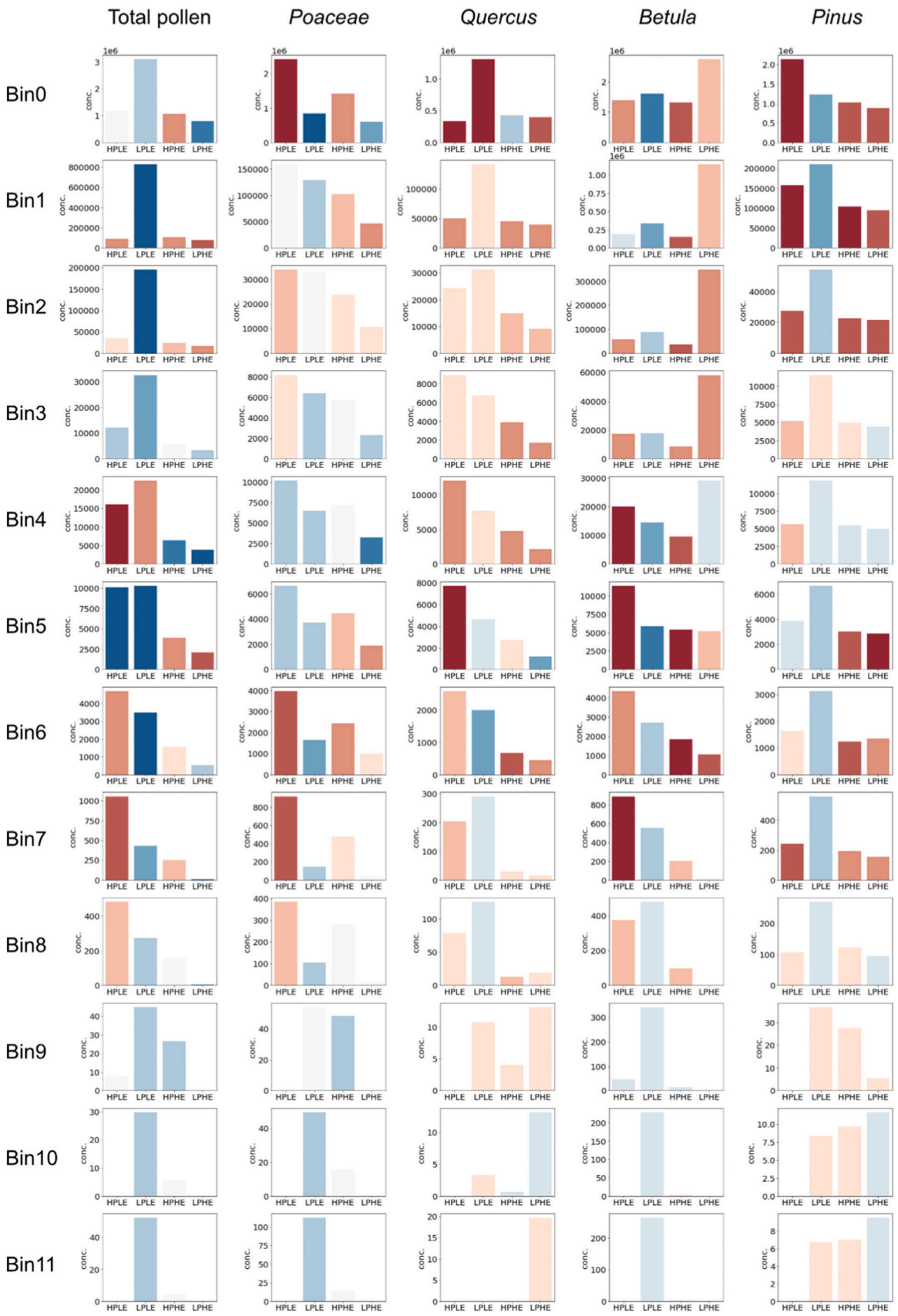


Fig. 4. Bar plots showing feature value for bins 0–11 for HPLE (high pollen low error), LPLE (low pollen high error), HPHE (high pollen high error) and LPHE (low pollen high error) events respectively for each target pollen variable. The colour of the bars represents an arbitrary scale for the corresponding SHAP value (i.e. impact on the model output), with a stronger red meaning the given feature value pushed the model output (pollen concentration) up and a stronger blue meaning vice versa, the model output was pulled down. The particle sizes sensed by each bin are tabulated in Table S1: the particle size range increases with bin number.

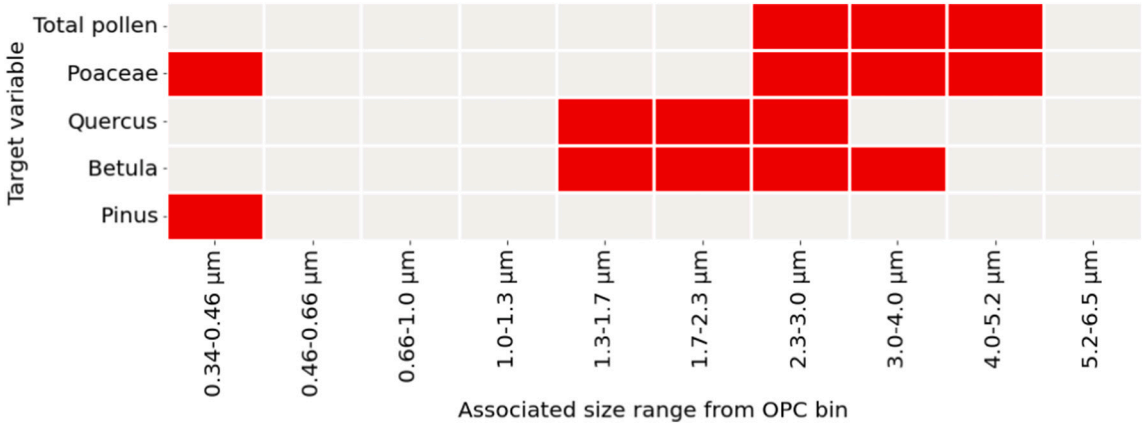


Fig. 5. Summary of particle sizes from OPC bins that demonstrate some positive correlation (shaded red) with each target variable according to further SHAP analysis (including the information from bar plots in Fig. 4). Note: The different bins/cells correlate positively to varying extents (e.g. 1.7–2.3 μm has relatively stronger correlation than adjacent bins for Quercus) but this is difficult to quantify and is not taken into account in this figure.

be put to purpose, for example among pollen monitoring networks for public health, it may be necessary to produce trained models for a particular context and scale within which there is some consistency of the local pollen environment. For a model to be used across a large scale among different contexts, i.e. different regions or countries, it would likely need a lot of varied data from within these regions for training.

Nevertheless, we have also demonstrated the effect that methodical hyperparameter tuning can have on increasing model performance and that this works best on a case-by-case basis, depending on the task and available data. The method we have presented in this study could be further applied to optimise models that take fewer or different input features, which may be able to achieve even better results.

4. Conclusions

In this study, we have demonstrated how the methodology to detect pollen using low-cost OPC sensors presented previously in Mills et al. (2023) can be improved by hyperparameter tuning and regularisation techniques for better performance. We have also demonstrated its application on an increased range of pollen types. For hourly and daily time resolution models, 6/10 and 8/10 respectively were able to produce R^2 scores above 0.5 for the test dataset. Maximum R^2 scores achieved were 0.89 and 0.95 for hourly and daily models respectively, once again for the models targeting Poaceae pollen with particle size and meteorological input data.

Using the SHAP explainable AI (XAI) method, we visualised and described the observable relationships between input features and target pollen concentration and assessed which particle size ranges were responsible for pushing up or pulling down predicted outputs. We further investigated the relationship between specific bin particle concentrations and impact on model output under four selected circumstances of high/low pollen and high/low error respectively. In particular, *Quercus* pollen displayed strong positive correlation with particles in the range 1.3–3.0 μm, which could be evidence of associated subpollen particles in this size range. *Quercus* and *Betula* pollen showed some similarity with each other but difference from *Poaceae* (which largely contributed to total pollen), while *Pinus* pollen showed mild similarity with *Poaceae*. Thus, this method may be more effective at isolating some types (e.g. *Quercus*) than others, and at differentiating between certain types (e.g. *Quercus* vs *Poaceae*) based on particle size measurements alone.

This work demonstrates the potential that this method shows for low-cost monitoring of pollen and provides deeper explanation as to how the models are learning and the influence each input feature has on predicting output pollen concentrations. While caution must always be taken when implying causal inference from machine learning models, useful information is obtainable that we can pair with our current scientific understanding. Currently, our scientific understanding of subpollen particles is limited and further studies characterising these particles for different pollen types would be an important step forward to provide scientific verification for these methods. Meanwhile, as with most machine learning methods, this technique would greatly benefit from training and testing on more data representing more varied environments to truly assess the extent of generalisability.

Funding

We acknowledge the funding support of the Natural Environment Research Council (NERC) CENTA2 grant NE/S007350/1 via the University of Birmingham, and the grant "Quantification of Utility of Atmospheric Network Technologies (QUANT)" (NE/T001968/1). The intercomparison campaign where data was obtained for this study was funded by the Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit (LGL) and EUMETNET AutoPollen Programme. Financial support was also received for this from the COST Action CA18226 ADOPT – *New approaches in detection of pathogens and aeroallergens*.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data supporting this publication are openly available from the UBIRA eData repository at doi: https://doi.org/10.25500/edata.bham.00000898.

Acknowledgements

We thank Andrew Tanner from the University of Birmingham, UK, for his technical work on functionalising the Alphasense OPC sensors for the purposes of this study. We thank the ZAUM (Jeroen Buters, Carsten B. Schmidt-Weber, Cordula Ebner von Eschenback, Gudrun Pusch, Marina Triviño and Christine Weil) and MeteoSwiss (Bernard Clot, Nina Burgdorfer and Sophie Erb) teams for their assistance facilitating the EUMETNET AutoPollen – COST ADOPT intercomparison campaign, as well as the Helmholtz Zentrum München as the campaign host and their assisting team (Daphne Kolland, George Matuscheck and Benjamin Schnautz)s. The manual pollen analysts for the campaign Łukasz

Kostecki and Agata Szymanska, from Łukasz Grewling's laboratory at Adam Mickiewicz University (Poznan, Poland) are greatly appreciated for their work facilitating the vital reference data for this campaign. We thank Robert Gebauer, Gisela Nagy and Anton Pointner for their IT support during the campaign.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2023.165853.

References

Adamov, S., Lemonis, N., Clot, B., Crouzy, B., Gehrig, R., Graber, M.-J., Sallin, C., Tummon, F., 2021. On the measurement uncertainty of Hirst-type volumetric pollen and spore samplers. Aerobiologia. https://doi.org/10.1007/s10453-021-09724-5.

Bacsi, A., Choudhury, B.K., Dharajiya, N., Sur, S., Boldogh, I., 2006. Subpollen particles: carriers of allergenic proteins and oxidases. J. Allergy Clin. Immunol. 118 (4), 844–850. https://doi.org/10.1016/j.jaci.2006.07.006.

Bengio, Y., 2012a. Practical recommendations for gradient-based training of deep architectures. In: Montavon, G., Orr, G.B., Müller, K.R. (Eds.), Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science, vol. 7700. Springer, Berlin, Heidelberg.

Bengio, Y., 2012b. Practical for gradient-based training of deep architectures. In: Neural Networks: Tricks of the Trade, Second edition, pp. 437–478. https://doi.org/10.48550/arXiv.1206.5533.

Bousiotis, D., Alconcel, L.N.S., Beddows, D.C., Harrison, R.M., Pope, F.D., 2023. Monitoring and apportioning sources of indoor air quality using low-cost particulate matter sensors. Environ. Int. 174, 107907 https://doi.org/10.1016/j.envint.2023.107907.

Bradley, R.S., 2015. Paleoclimatology: Chapter 12 – Pollen, 3rd ed. Elsevier, Massachusetts, USA, pp. 408–409. https://doi.org/10.1016/B978-0-12-386913-5.00012-0.

Brownlee, J., 2018. Better Deep Learning. Edition: v1.8. https://machinelearningmastery.com/better-deep-learning/.

Burkart, J., Gratzl, J., Seifried, T.M., Bieber, P., Grothe, H., 2021. Isolation of subpollen particles (SPPs) of birch: SPPs are potential carriers of ice nucleating macromolecules. Biogeosciences 18, 5751–5765. https://doi.org/10.5194/bg-18-5751-2021.

Buters, J.T.M., Antunes, C., Galveias, A., Bergmann, K.C., Thibaudon, M., Galán, C., Schmidt-Weber, C., Oteros, J., 2018. Pollen and spore monitoring in the world. Clin. Transl. Allerg. 8, 9. https://doi.org/10.1186/s13601-018-0197-8.

Buters, J., Clot, B., Galán, C., Gehrig, R., Gilge, S., Hentges, F., O'Connor, D., Sikoparija, B., Skjoth, C., Tummon, F., Adams-Groom, B., Antunes, C.A., Bruffaerts, N., Çelenk, S., Crouzy, B., Guillaud, G., Hajkova, L., Kofol Seliger, A., Oliver, G., Ribeiro, H., Rodinkova, V., Saarto, A., Sauliene, I., Sozinova, O., Stjepanovic, B., 2022. Automatic detection of airborne pollen: an overview. Aerobiologia. https://doi.org/10.1007/s10453-022-09750.

Chappuis, C., Tummon, F., Clot, B., Konzelmann, T., 2020. Automatic pollen monitoring: first insights from hourly data. Aerobiologia 36, 159–170. https://doi.org/10.1007/s10453-019-09619-6.

Christianini, N., 2010. Are we there yet? Neural Netw. 23 (4), 466–470. https://doi.org/10.1016/j.neunet.2010.01.006.

Crilley, L.R., Singh, A., Kramer, L.J., Shaw, M.D., Alam, M.S., Apte, J.S., Bloss, W.J., Hildebrandt Ruiz, L., Fu, P., Fu, W., Gani, S., 2020. Effect of aerosol composition on the performance of low-cost optical particle counter correction factors. Atmos. Meas. Techniq. 13 (3), 1181–1193.

Cristianini, N., 2023. The Shortcut: Why Intelligent Machines Do Not Think like us. CRC Press. ISBN 9781032305097.

Crouzy, B., Stella, M., Konzelmann, T., Calpini, B., Clot, B., 2016. All-optical automatic pollen identification: towards an operational system. Atmos. Environ. 140, 202–212. https://doi.org/10.1016/j.atmosenv.2016.05.062.

Darrow, L.A., Hess, J., Rogers, C.A., Tolbert, P.E., Klein, M., Sarnat, S.E., 2012. Ambient pollen concentrations and emergency department visits for asthma and wheeze. J. Allergy Clin. Immunol. 130 (3), 630–638. https://doi.org/10.1016/j.jaci.2012.06.020.

Després, V.R., Huffman, J.A., Burrows, S.M., Hoose, C., Safatov, A.S., Buryak, G., Fröhlich-Nowoisky, J., Elbert, W., Andreae, M.O., Pöschl, U., Jaenicke, R., 2012. Primary biological aerosol particles in the atmosphere: a review. Tellus B 64. https://doi.org/10.3402/tellusb.v64i0.15598.

Diehl, K., Quick, C., Matthias-Maser, S., Mitra, S.K., Jaenicke, R., 2001. The ice nucleating ability of pollen: part I: laboratory studies in deposition and condensation freezing modes. Atmos. Res. 58 (2), 75–87. https://doi.org/10.1016/S0169-8095(01)00091-6.

Diehl, K., Matthias-Maser, S., Jaenicke, R., Mitra, S.K., 2002. The ice nucleating ability of pollen: part II: laboratory studies in immersion and contact freezing modes. Atmos. Res. 61 (2), 125–133. https://doi.org/10.1016/S0169-8095(01)00132-6.

Dreischmeier, K., Budke, C., Wiehemeier, L., Kottke, T., Koop, T., 2017. Boreal pollen contain ice-nucleating as well as ice-binding 'antifreeze' polysaccharides. Sci. Rep. 7, 41890. https://doi.org/10.1038/srep41890.

Fröhlich-Nowoisky, J., Kampf, C.J., Weber, B., Huffman, J.A., Pöhlker, C., Andreae, M.O., Lang-Yona, N., Burrows, S.M., Gunthe, S.S., Elbert, W., Su, H., Hoor, P., Thines, E., Hoffmann, T., Després, V.R., Pöschl, U., 2016. Bioaerosols in the earth system: climate, health, and ecosystem interactions. Atmos. Res. 182, 346–376. https://doi.org/10.1016/j.atmosres.2016.07.018.

Gohel, P., Singh, P., Mohanty, M., 2021. Explainable AI: current status and future directions. IEEE Access. https://doi.org/10.48550/arXiv.2107.07045 arXiv: 2107.07045.

Goodfellow, I., Bengio, Y., Courville, A., 2017. Deep Learning. The MIT Press, Cambridge. ISBN: 9780262035613.

Griffiths, P.T., Borlace, J.-S., Gallimore, P.J., Kalberer, M., Herzog, M., Pope, F.D., 2012. Hygroscopic growth and cloud activation of pollen: a laboratory and modelling study. Atmos. Sci. Lett. 13 (4), 289–295. https://doi.org/10.1002/asl.397.

Gute, E., Abbatt, J.P.D., 2020. Ice nucleating behaviour of different tree pollen in the immersion mode. Atmos. Environ. 231, 117488 https://doi.org/10.1016/j.atmosenv.2020.117488.

Hendrickson, B.N., Alsante, A.N., Brooks, S.D., 2023. Live oak pollen as a source of atmospheric particles. Aerobiologia 39, 51–67. https://doi.org/10.1007/s10453-022-09773-4.

Huffman, J.A., Perring, A.E., Savage, N.J., Clot, B., Crouzy, B., Tummon, F., Shoshanim, O., Damit, B., Schneider, J., Sivaprakasam, V., Zawadowicz, M.A., Crawford, I., Gallagher, M., Topping, D., Doughty, D.C., Hill, S.C., Pan, Y., 2019. Real-time sensing of bioaerosols: review and current perspectives. Aerosol Sci. Technol. 54 (5), 465–495. https://doi.org/10.1080/02786826.2019.1664724.

Jiang, C., Wang, W., Du, L., Huang, G., McConaghy, C., Fineman, S., Liu, Y., 2022. Field evaluation of an automated pollen sensor. Int. J. Environ. Res. Public Health 19 (11), 6444. https://doi.org/10.3390/ijerph19116444.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Proces. Syst. 30, 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mäkelä, E.M., 2009. Size distinctions between Betula pollen types – a review. Grana 35 (4), 248–256. https://doi.org/10.1080/00173139609430011.

Marchisio, A., Hanif, M.A., Rehman, S., Shafique, M., 2018. A methodology for automatic selection of activation functions to design hybrid deep neural networks, arXiv: 1811.03980v1. doi:10.48550/arXiv.1811.03980.

Masters, D., Luschi, C., 2018. Revisiting Small Batch Training for Deep Neural Networks. arXiv:1804.07612. doi:10.48550/arXiv.1804.07612.

Matthews, B.H., Alsante, A.N., Brooks, S.D., 2023. Pollen emissions of subpollen particles and ice nucleating particles. ACS Earth Space Chem. 7 (6), 1207–1218. https://doi.org/10.1021/acsearthspacechem.3c00014.

Maya-Manzano, J.M., Tummon, F., Abt, R., Allan, N., Bunderson, L., Clot, B., Crouzy, B., Daunys, G., Erb, S., Gonzalez-Alonso, M., Graf, E., Grewling, Ł., Haus, J., Kadantsev, E., Kawashima, S., Martinez-Bracero, M., Matavilj, P., Mills, S., Niederberger, E., Lieberherr, G., Lucas, R.W., O'Connor, D.J., Oteros, J., Palamarchuk, J., Pope, F.D., Rojo, J., Šaulienė, I., Schäfer, S., Schmidt-Weber, C.B., Schnitzler, M., Šikoparija, B., Skjøth, C.A., Sofiev, M., Stemmler, T., Triviño, M., Zeder, Y., Buters, J., 2023. Towards European automatic bioaerosol monitoring: comparison of 9 automatic pollen observational instruments with classic Hirst-type traps. Sci. Total Environ. 866, 161220 https://doi.org/10.1016/j.scitotenv.2022.161220.

McCurry, J., 2023. "Japan declares war on pollen as hay fever epidemic grips the nation", The Guardian. https://www.theguardian.com/world/2023/apr/06/japan-declares-war-on-pollen-as-hay-fever-epidemic-grips-the-nation, (accessed 10/05/2023).

Mikhailov, E.F., Ivanova, O.A., Nebosko, E.Y., Vlasenko, S.S., Ryshkevich, T.I., 2019. Subpollen particles as atmospheric cloud condensation nuclei. Izv. Atmos. Ocean. Phys. 55, 357–364. https://doi.org/10.1134/S000143381904008X.

Mills, S.A., Bousiotis, D., Maya-Manzano, J.M., Tummon, F., MacKenzie, A.R., Pope, F.D., 2023. Constructing a pollen proxy from low-cost optical particle counter (OPC) data processed with neural networks and random forests. Sci. Total Environ. 871, 161969 https://doi.org/10.1016/j.scitotenv.2023.161969.

Oteros, J., Pusch, G., Weichenmeier, I., Heimann, U., Möller, R., Röseler, S., Traidl-Hoffmann, C., Schmidt-Weber, C., Buters, J.T.M., 2015. Automatic and online pollen monitoring. Int. Arch. Allergy Immunol. 167, 158–166. https://doi.org/10.1159/000436968.

Oteros, J., Weber, A., Kutzora, S., Rojo, J., Heinze, S., Herr, C., Gebauer, R., Schmidt-Weber, C.B., Buters, J.T.M., 2020. An operational robotic pollen monitoring network based on automatic image recognition. Environ. Res. 191, 110031 https://doi.org/10.1016/j.envres.2020.110031.

Pedamonti, D., 2018. Comparison of non-linear activation functions for deep neural networks on MNIST classification task. arXiv:1804.02763. doi:10.48550/arXiv.1804.02763.

Pichler, M., Hartig, F., 2023. Machine learning and deep learning – a review for ecologists. Methods Ecol. Evol. 14 (4), 994–1016. https://doi.org/10.1111/2041-210X.14061.

Pope, F.D., 2010. Pollen grains are efficient cloud condensation nuclei. Environ. Res. Lett. 5, 044015 https://doi.org/10.1088/1748-9326/5/4/044015.

Pummer, B.G., Bauer, H., Bernardi, J., Bleicher, S., Grothe, H., 2012. Suspendable macromolecules are responsible for ice nucleation activity of birch and conifer pollen. Atmos. Chem. Phys. 12, 2541–2550. https://doi.org/10.5194/acp-12-2541-2012.

Radaeski, J.N., Bauermann, S.G., Pereira, A.B., 2016. Poaceae pollen from southern Brazil: distinguishing grasslands (Campos) from forests by Analysing a diverse range of Poaceae species. Front. Plant Sci. 7 doi:10.3389%2Ffpls.2016.01833.

Reed, R., Marks, R.J., 1999. Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks. The MIT Press, ISBN 9780262527019.

Reponen, T., 2011. Encyclopedia of Environmental Health: Methodologies for Assessing Bioaerosol Exposures. Elsevier, Cincinnati, USA, p. 723. https://doi.org/10.1016/B978-0-12-409548-9.11822-6.

Šaulienė, I., Šukienė, L., Daunys, G., Valiulis, G., Vaitkevičius, L., Matavulj, P., Brdar, S., Panic, M., Sikoparija, B., Clot, B., Crouzy, B., Sofiev, M., 2019. Automatic pollen recognition with the rapid-E particle counter: the first-level procedure, experience and next steps. Atmos. Meas. Tech. 12 (6), 3435–3452. https://doi.org/10.5194/amt-12-3435-2019.

Sauvageat, E., Zeder, Y., Auderset, K., Calpini, B., Clot, B., Crouzy, B., Konzelmann, T., Lieberherr, G., Tummon, F., Vasilatou, K., 2020. Real-time pollen monitoring using digital holography. Atmos. Meas. Tech. 13, 1539–1550. https://doi.org/10.5194/amt-13-1539-2020.

Smiljanic, K., Apostolovic, D., Trifunovic, S., Ognjenovic, J., Perusko, M., Mihajlovic, L., Burazer, L., van Hage, M., Cirkovic Velickovic, T., 2017. Subpollen particles are rich carriers of major short ragweed allergens and NADH dehydrogenases: quantitative proteomic and allergomic study. Clin. Exp. Allergy 47 (6), 815–828. https://doi.org/10.1111/cea.12874.

Song, U., Park, J., Song, M., 2012. Pollen morphology of Pinus (Pinaceae) in northeast China. For. Sci. Technol. 8 (4), 179–186. https://doi.org/10.1080/21580103.2012.704973.

Steiner, A.L., Brooks, S.D., Deng, C., Thornton, C.O., Pendleton, M.W., Bryant, V., 2015. Pollen as atmospheric cloud condensation nuclei. Geophys. Res. Lett. 42 (9), 3596–3602. https://doi.org/10.1002/2015GL064060.

Stone, E.A., Mampage, C.B.A., Hughes, D.D., Jones, L.M., 2021. Airborne sub-pollen particles from rupturing giant ragweed pollen. Aerobiologia 37, 625–632. https://doi.org/10.1007/s10453-021-09702-x.

Subba, T., Zhang, Y., Steiner, A.L., 2023. Simulating the transport and rupture of pollen in the atmosphere. J. Adv. Model. Earth Syst. 15 (3) https://doi.org/10.1029/2022MS003329 e2022MS003329.

Tong, H.-J., Ouyang, B., Nikolovski, N., Lienhard, D.M., Pope, F.D., Kalberer, M., 2015. A new electrodynamic balance (EDB) design for low-temperature studies: applicationto immersion freezing of pollen extract bioaerosols. Atmos. Meas. Tech. 8, 1183–1195. https://doi.org/10.5194/amt-8-1183-2015.

Triviño, M.M., Maya-Manzano, J.M., Tummon, F., Clot, B., Grewling, Ł., Schmidt-Weber, C., Buters, J., 2023. Variability between Hirst-type pollen traps is reduced by resistance-free flow adjustment. Aerobiologia 39, 257–273. https://doi.org/10.1007/s10453-023-09790-x.

Wrońska-Pilarek, D., Danielewicz, W., Bocianowski, J., Malinski, T., Janyszek, M., 2016. Comparing pollen morphological analysis and its systematic implications on three European oak (Quercus L., Fagaceae) species and their spontaneous hybrids. PLoS One 11 (8), e0161762. https://doi.org/10.1371/journal.pone.0161762.