Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 625–630 DOI: 10.15439/2024F8703 ISSN 2300-5963 ACSIS, Vol. 39

Plant-traits: how citizen science and artificial intelligence can impact natural science

Giacomo Ignesti
0000-0003-2389-3086
CNR-ISTI
Via Giuseppe Moruzzi, 1, 56124 Pisa, Italy
University of Pisa
Second Floor, Largo Bruno Pontecorvo, 3, 56127 Pisa, Italy
Email: giacomo.ignesti@isti.cnr.it

Davide Moroni, Massimo Martinelli
0000-0002-5175-5126
0000-0001-7419-5099
CNR-ISTI
Via Giuseppe Moruzzi, 1, 56124 Pisa, Italy
Email: {name.surname}@isti.cnr.it

***Abstract*—Citizen science has emerged as a valuable resource for scientific research, providing large volumes of data for training deep learning models. However, the quality and accuracy of crowd-sourced data pose significant challenges for supervised learning tasks such as plant trait detection. This study investigates the application of AI techniques to address these issues within natural science. We explore the potential of multi-modal data analysis and ensemble methods to improve the accuracy of plant trait classification using citizen science data. Additionally, we examine the effectiveness of transfer learning from authoritative datasets like PlantVillage to enhance model performance on open-access platforms such as iNaturalist. By analysing the strengths and limitations of AI-driven approaches in this context, we aim to contribute to developing robust and reliable methods for utilising citizen science data in natural science.**

I. Introduction

CITIZEN SCIENCE (CS) is a valuable approach involving the public in scientific research activities [1], [2]. The trade-offs of this approach are well known; while it generates a vast amount of data and fosters public trust in science, data quality may vary due to the different levels of expertise among participants. The construction and maintenance of a CS dataset are important topics that deserve to be treated separately. Still, it is equally essential to analyse the collected data: the uncertainty of these collections must be treated adequately. Modern Machine Learning (ML) and Deep Learning (DL) models can help since these algorithms are structured to automatically process large amounts of data and show partial resilience to the collected data's precision problem. CS projects are mostly related to natural sciences; in this domain, two of the most active Websites and databases are eBird [3] for ornithology and iNaturalist [4] for capturing images of the natural world. Beyond these, several examples of online communities, like those on Zooniverse [5], serve as significant, multifaceted incubators for scientific discovery. A notable example of the successful integration of DL and CS in natural science is represented by the study by Schiller et al. [6] that demonstrates the potential to automate plant traits predictions from photographs. In the last few years, different other CS datasets have influenced prominent work [7] with a focus on how to interpret and use CS data correctly. Still, Schiller's seminal work has highly impacted DL and natural science, inspiring subsequent study [8]. Schiller et al. exhibit the potential of multi-modal DL models using a smart combination of picture and tabular bioclimatic data with a multi-step pipeline that encompasses and ensemble three baseline models to fit plant traits, mainly a CNN network, a trait variability informed network, and an ANN for the tabular data. They achieved this result by integrating information from three key sources: i.e. Citizen science iNaturalist image repositories, the TRY dataset [9] that contains plant traits, and the WorldClim [10] dataset for bioclimatic data. The brilliance of their approach lies in leveraging the image labels (species and geolocation data) of the iNaturalist dataset. These labels serve as unique identifiers, allowing researchers to integrate plant image data with trait and climate information stored in the two high-level scientific repositories. Since its release in late 2021, their article has directly inspired two worldwide AI competitions sponsored by the Fine-Grained Visual Categorisation (FGVC) workshop at the Conference on Computer Vision and Pattern Recognition (CVPR), one of the most relevant conferences for DL methods applied to images and signals. The competitions available on the Kaggle Website are *PlantTraits2023* [11] and *PlantTraits2024* [12]. This paper briefly explores how AI can significantly enhance citizen science research. Starting from the plant-traits related dataset and their objective, it is evaluated how the performance of different ML and DL algorithms change between these repositories. Inconsistencies and similarities between the results can be used to assess the relation between data quality, model and task to build a starting guide for researchers to start or improve their work. The three datasets are initially evaluated with standard tabular data ML analysis (i.e. XGBoost and catboost) and then further studied with DL solution to process image and other data types together, starting from the *PlantTraits2023* winner solution. To this end, the reported study further expanded the studies of plant automatic processing initiated in [13] and assessed how winning strategies as efficient adaptive ensemble and transfer learning for plant classification algorithms perform using these three different datasets of similar domains (plants). It also presented an adaptive ensemble of ConvNeXt-V2 [14] and a classification to regression transfer learning strategy to lay the groundwork for developing robust AI tools in a natural science framework. The study is presented in the following sections. In Section II, materials and methods are introduced, detailing the three datasets' structure and analysis and describing the ML and DL models used to process the contained data. Section III presents results, introducing metrics and reporting the models' performance to provide a comprehensive overview of the outcomes. Section IV is dedicated to the discussion, delving into the insights gained from the results and exploring the implications and limitations of the findings. The study's key takeaways, potential highlights, and future work are presented in the conclusions (Section V).

II. Materials and Methods

A. Dataset

The investigation has been conducted on three available open-access datasets:

- the first one, named *PlantTraits2021* (`PT2021`), is accessible from the article repository [15], where it is explained how to merge the different files to obtain the complete dataset; to date, the link of each image is deprecated, but the tabular data is still accessible to analyse their distribution;
- the second *PlantTraits2023* (`PT2023`) repository is available under the Kaggle competition having the same name: both image and tabular data are available;
- the third *PlantTraits2024* (`PT2024`) is also available under the Kaggle competition having the same name: both image and tabular data are available.

The three datasets were curated to predict the plant traits, which are the target variables to obtain throughout a regression approach: while `PT2021` and `PT2024` contain the same six target traits, `PT2023` contain 34 target traits. The three datasets share common possible targets, enabling comparisons between them. The six shared targets are some of the most valuable plant traits [6]: stem specific density (SSD), leaf area per leaf dry mass, plant height, seed dry mass, leaf nitrogen (N) content per leaf area, and leaf area. As already stated, the input information is composed of RGB images and tabular data: the tabular ones are mainly numerical, except for the species and geolocation variables; the species variable is standard to the `PT2021` and `PT2024` versions, while the geolocation attribute is present only in the `PT2021` version of the dataset. The three datasets share a baseline of four bio-climatic information, noted in the literature as bioX. Bio-climatic variables represent climatic data, such as temperature and precipitation, with the perspective of their influence on the biological sphere. The four variables and the six common plant traits targets are used to build the three datasets reported in the study as the *minimal informed datasets*. These datasets are essentially the minor subsets of the original datasets obtained by removing columns while containing the maximum number of shared data Table I. Beyond this baseline, the `PT2021` and `PT2023` datasets primarily contain bio-climatic variables. In contrast, the `PT2024` dataset expands the input variable space by incorporating multi-temporal data from satellite sources, including Moderate Resolution Imaging Spectroradiometer (MODIS), which provides near-real-time Earth surface reflection information, radar data from the Vegetation Optical Depth (VOD) dataset for measuring vegetation density, and soil information indicating key components for plant growth, such as nutrients or moisture content.

TABLE I
Analysed Datasets: Original and Synthesised with number of features [feat] and target [tgt]

Dataset	Original	Minimal Shared	Max Size
`PT2021`	25 feat. | 6 tgt.	4 feat. | 6 tgt.	43,745,637
`PT2023`	18 feat. | 34 tgt.	4 feat. | 6 tgt.	1,921,780
`PT2024`	163 feat. | 6 tgt.	4 feat. | 6 tgt.	9,766,064

B. Methods

1) Preprocessing: The first analysis performed was an evaluation of the percentage of missing data in each dataset; after that, all data outside the 95% interquartile range were removed. Further cleaning was then performed: specifically, all data violating physical limits, such as negative absolute percentages or extensive values (self-evidently, lengths unit can not be negative), were removed from the remaining data. Lastly, the target plant traits and shared tabular input are evaluated to detect if data follow the same distribution and value range; unit measure should be coherent between the three datasets even if there are no explicit units for `PT2023` and `PT2024`.

2) ML Model: Once the three sets of data are obtained, ML and DL algorithms are fitted to them; in particular, ML models are a vital ingredient for comparison since they can be used as a baseline for the three datasets, as tabular data is shared between datasets. Two different ML models were trained over the three completed datasets and their *minimal informed datasets* versions; the objective is to estimate the plant traits. These are the models used: eXtreme Gradient Boosting (XGBoost) [16] and CatBoost [17]. The general workflow is shown in Figure II-B2. Each model was evaluated using k-fold cross-validation, where the data is split into k subsets. The model is trained k times, using $k-1$ subsets for training and the remaining subset for testing each iteration. This study used $k=5$, resulting in an 80/20 train-validation split since, technically, the two other datasets can be used as a separate test set. A comparison between a model operating on the *full input features* and one on the *minimal shared features* is computed.

3) DL Model: The second experiment was set to reproduce the results of the winner of the `PT2023` competition, firstly on the same year dataset and then on the `PT2024` one. The `PT2023` competition solution leveraged a large-scale Fully Convolutional Masked Autoencoder (FCMAE) from the ConvNeXt V2 family as the image processing backbone. A novel approach to plant trait evaluation complemented this state-of-the-art architecture. Rather than treating the 13 plant

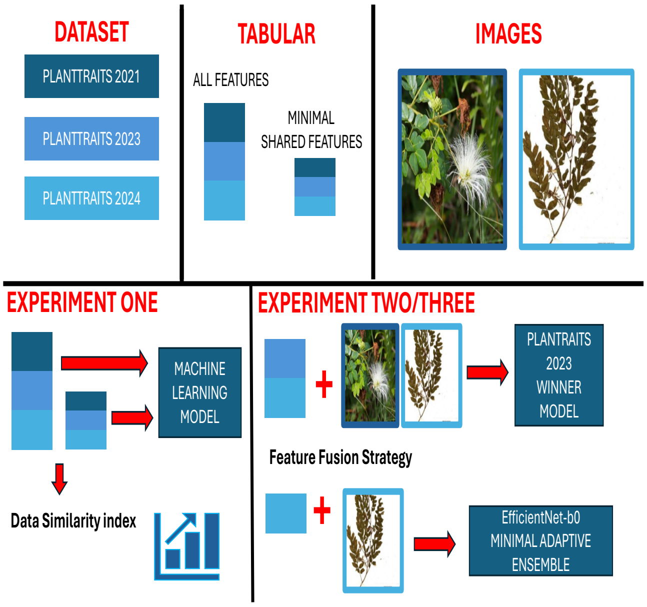

Fig. 1. Study design and execution

traits as regression targets, they were mapped to class labels, transforming the problem into a classification task. Other than standard data augmentation operation on the whole dataset [Random Resized Crop, Transpose, Horizontal Flip, Vertical Flip, Piece-wise Affine, Hue Saturation Value, Random Brightness Contrast], while a CUT-MIX [18] data augmentation operation is randomly applied on a portion of the training set. As a feature fusion solution, the backbone's output was concatenated with a tensorized version of input metadata before being fed into a fully connected layer to produce a 12512-dimensional output. Two models were trained using this approach, differing in the application of CUT-MIX across a percentage of the training set [90%-80%] and in the dropout values applied after the stacking layer [0.50-0.68]; the outputs of the models are then ensembled with a bagging technique without voting. Following the approaches described in [19], the minimal adaptive ensemble was estimated to train together the two models to classify the target class. The operation has then repeated for the `PT2024`. All other relevant information can be found in the archive of the code repository[1]. The last experiment was meant to work as a solution to the `PT2024` competition. A pre-trained EfficientNet-b0 model was employed for transfer learning; pre-training was conducted on the PlantVillage dataset [20] that was chosen because of domain similarity. The fully connected (FC) layer of the EfficientNet-b0 architecture was removed since the original network was trained for plant classification and not for regression. The best-obtained model is then combined in a minimal adaptive ensemble, as previously explained, and the resulting model is used to fit the six regression values(numerical target). This process was repeated by training the EfficientNet-b0 model

[1]https://github.com/DuanChenL/FGVC10

TABLE II
NORMALISING VALUE USED IN THE DL EXPERIMENT

DATASET	MEAN	STD
2024	[0.3356, 0.4496, 0.4446]	[0.2355, 0.2260, 0.2348]
2023	[0.3356, 0.4580, 0.4398]	[0.2376, 0.2281, 0.2360]
ImageNet1k	[0.485, 0.456, 0.406]	[0.229, 0.224, 0.225]
PLANTRAITS	[0.5258, 0.5357, 0.5277]	[0.1530 0.1249, 0.1142]

without freezing weights (full training). The resulting models were then combined into an ensemble for comparison. The best ensemble is then updated with the features of tabular data with a feature fusion approach. The tabular data are inputted inside an FC layer for processing and then concatenated to the feature of the ensemble; an FC layer then processes the fused features. A slight variation of this structure was also proposed and consists of passing the tabular data to an FT-Transformer [21]. In this case, features are passed to a numerical embedding layer and then to the multi-head attention structure of the transformer. Such structure outputs a high-dimensional embedded vector fused with the ensemble model features and processed following the already described procedure. In summary, for the last experiment on the `PT2024`, the tested architectures are i) the EfficientNet-b0 classic CNN structure, ii) the minimal adaptive ensemble of the architecture, iii) the informed ensemble architecture, and iv) the minimally informed ensemble architecture. The number of epochs used for training is variable since it is used as early-stopping criteria to avoid overfitting. Other particular settings used to train the network are an image size of 224x224 and a one-cycle learning rate policy. Images were normalised using the original dataset's plant-trait values in the transfer learning setting (Table II). In contrast, the values for the comparison model were normalised using their original variance. All the experiments were performed using Python and two devices: an NVIDIA GeForce RTX 4060 and a pair of NVIDIA Quadro RTX 5000.

III. RESULTS

The percentage of missing data among the three datasets is coherent at around 1% in `PT2021` and `PT2024`, while no missing entries are present in the `PT2023` dataset. In both cases, missing data is associated with the column containing the traits' standard deviation values, with an average of 30% missing data per column. These six datasets (three original and three minimal) are then analysed to identify outliers (data points significantly different from the rest) and ensure the remaining data is physically coherent (meaningful and consistent). This operation removes around 15% of the data in each case. The last check on tabular data dimension is on the distribution of the standard input features and the outcome. Regarding input features, the distribution along the dataset is similar, with a slight difference in the range. The distribution of plant traits, referring to the six shared columns, appears that the `PT2024` and `PT2023` datasets follow the same distribution, while `PT2021` seems to contain more sparse data FigIIIIII. The image training set normalisation values are very

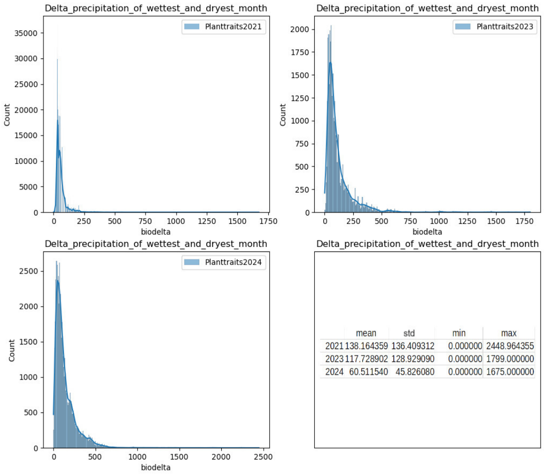

Fig. 2. Distribution "Delta of the precipitation of wettest and dryest months", a common feature across all the three datasets, no unit measure used since the lack in two of the three datasets

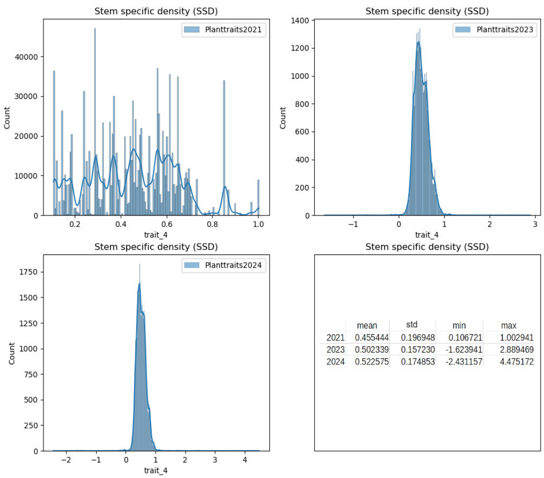

Fig. 3. Distribution "Stem-specific density", a common feature across all the three datasets, no unit measure used since the lack in two of the three datasets

similar among the `PT2023` and `PT2024` datasets.

A. *ML Results*

The performance of catboost and xgboost algorithms along the three datasets is similar. Indeed, the mean square error magnitude and the mean absolute error magnitude maintain the same order between the three original datasets and the minimal composed dataset. Training the algorithms over the whole dataset tends to output higher performances in all analysed cases. The `PT2024` dataset seems to contain the most difficult task per number of elements while training with all the feature boost accuracy in the `PT2023` and `PT2021` case

B. *DL Results*

Using the 2023 competition winner solution gives mixed results. The R2 score of the reproduced model [69%] on the same-year repository test set is similar to that reported by the winner [72%]. The adaptive ensemble tested shows overall the same accuracy, obtaining [68%] accuracy improving from the base original two models by [2%] and [6%]. Extensive experiment was conducted to reproduce these results on the `PT2024` with no success. The network shows an accuracy of classification of [84%] during training, but a negative score on the R2 was achieved on the test set. The successive experiment adopted to solve the `PT2024` challenge with transfer learning also does not seem to perform well, even achieving better performance than the 2023 model solution. The maximum accuracy on the related minimum weak is around the R2 SCORE value of 0.1 on the validation set and less than 0.1 on the separate test set; the ensemble of the pre-trained weak learner does not boost the overall accuracy. Training the network from scratch using the minimal adaptive ensemble generates slightly better results, moving the R2 score towards 0.15 on the separate test set and around 0.45 on the validation set. The feature fusion approach is the one that obtains the higher accuracy, both on the validation set and on the separated test set. The performance of the model trained on the original 2024 dataset is slightly superior, 0.3 higher, concerning the model trained on the minimal shared data TableVI. In the two experiments using feature fusion with tabular data, there are only minimal differences between the approaches using classical ANN architectures for the TF-Transformer approach. Lastly, since the minimal informed dataset was trained with the compatible set of features and image types of the 2023 dataset, this is used as input to assert the inference. Still, the R2 metric is negative, so the model trained on the 2024 dataset seems inefficient in predicting traits using the information in the 2023 dataset.

IV. Discussion

Analysing these CS datasets through missing value counts, value distribution, and quartile ranges reveals their underlying relationships. The low percentage of missing data is related to the dataset construction criteria and the iNaturalist repository. iNaturalist boasts a 95% trust rating for Research-grade data; human error or inconsistencies in data collection can still occur, but replicating the criteria of Schiller et al., research should grant a coherent CS dataset. The estimated percentage of outlier quantity should not be accounted as an inconsistency but should be seen as a lack of total domain compression; some values as negative extensive measure unit or out of scale plant dimension Fig.IV are easy to detect, but outlier born from statistical anomaly is usually hard to detect. These findings justify the choice of a 95% interquartile range threshold; without extensive knowledge of data domain and source, this operation ensures a more controlled dataset concerning

TABLE III
CATBOOST TRAINING PERFORMANCE ON THE THREE DATASETS IN THE FIVE-FOLD CV SPLITS

Metric	catboost-MSE	catboost-MAE	catboost-R2score
Original 2024 dataset	[347374, 348317, 323868, 349014, 348052]	[162, 164, 159, 165, 163]	[0.19, 0.19, 0.18, 0.19, 0.20]
Minimal 2024 dataset	[3518134, 3461288, 3517697, 3495179, 3501414	[568, 565, 568, 568, 567]	[0.10, 0.09, 0.09, 0.09, 0.10]
Original 2023 dataset	[327861, 240540, 231010, 365676, 234271]	[130, 116, 112, 131, 114]]	[0.54, 0.64, 0.66, 0.53, 0.65]
Minimal 2023 dataset	[813215, 760145, 804344, 839923, 799965]	[239, 233, 235, 243, 237]	[0.15, 0.15, 0.15, 0.16, 0.1]
Original 2021 dataset	[85042, 78318, 81224, 137029, 78496]	[51, 49, 50, 54, 51]	[0.97, 0.97, 0.97, 0.97, 0.970]
Minimal 2021 dataset	[474667, 489234, 497001, 469697, 464775]	[187,192,189, 186,185]	[0.16, 0.15, 0.16, 0.15, 0.17]

TABLE IV
XGBOOST PERFORMANCE ON THE THREE DATASETS IN THE FIVE-FOLD CV SPLITS

Metric	XGBboost-MSE	XGBboost-MAE	XGBboost-R2score
Original 2024 dataset	[386514, 386673, 364084, 383395, 380447]	[168, 168, 163, 169, 166]]	[0.09, 0.10, 0.07, 0.08, 0.10]
Minimal 2024 dataset	[813215, 760145, 804344, 839923, 799965]	[557, 554, 557, 557, 55]	[0.13, 0.13, 0.13, 0.13 0.13]
Original 2023 dataset	[167278, 174812, 169189, 208440, 16210]	[48, 47, 50, 52, 49]	[0.75, 0.76, 0.73, 0.70, 0.75]
Minimal 2023 dataset	[920563, 870225, 926290, 919225, 910894]	[246, 240, 244, 246, 246]	[0.05, 0.05, 0.04, 0.07, 0.05]
Original 2021 dataset	[85042, 78318, 81224, 137029, 78496]	[1662, 59, 57, 80, 23]	[0.99, 0.99, 0.99, 0.99, 0.99]
Minimal 2021 dataset	[524289, 525170, 553655, 524170, 505197]	[192, 193, 193, 190, 189]]	[0.1, 0.1, 0.1, 0.08, 0.1]

TABLE V
DL MODEL RESULTS, THE R2 METRIC FOR PERFORMANCE EVALUATION DL MODEL RESULTS OF THE 2023 WINNING MODEL AND ITS ADJUSTMENT FOR THE PT2024 CHALLENGE AND THE ADAPTIVE ENSEMBLE

Model	Classification Accuracy [%]	Test R2 Metric
2023 Model 1	0.81	0.66
2023 Model 2	0.81	0.62
2023 Model Bagging E	0.83	0.69
2023 Model Adaptive E	0.83	0.68
2024 Model 1	0.81	-39

TABLE VI
DL MODEL RESULTS, THE R2 METRIC FOR PERFORMANCE EVALUATION

Model	Validation R2 Metric	Test R2 Metric
TF weak	0.10	0.08
TF Ensemble	0.15	0.10
Trained Weak	0.40	0.15
Trained Ensemble	0.53	0.18
I-Ensemble ANN	0.56	0.247
I-Ensemble ANN (2023)	//	-0.08
I-Ensemble FT-Transformer	0.56	0.248
Minimal I-Ensemble ANN	0.45	0.21

similar threshold values like 98% IQR. Nevertheless, the overall quality of these CS datasets is high, but a double check from a professional should always be done [22]. For future CS-based projects, it should be considered to implement data input controls at the point of entry. These controls can define plausible ranges for each data point and notify users of potential anomalies (out-of-range values) requiring validation. For this reason, the value distribution difference between the 2021 dataset and its counterpart is easily explicable. PT2021 contains more data, which is reflected in the fact that there are more plant species; some plants can be very different from each other Fig.III, so a much dense sample should result in a wider statistical distribution of the traits. Knowing that can help us detect what part of the dataset should be used in future work. The analysis of the 2024 dataset shows that the

Fig. 4. Example of an image of a plant with a presumed out-of-bound value: the leaf area is over two thousand squared meters

problem posed in this version of the competition appears more complex. The low percentage accuracy in training and separate testing sets lets us understand that the proposed ML model works correctly but fails to learn. In contrast, the accuracy obtained with the same article on the other two datasets is higher. The fact that the results between the 2024 minimal shared data information and the complete 2024 counterpart are very similar confirms the complexity of the task. It teaches us the possibility of using low-dimensional datasets for complex problems. In the 2021 and 2023 related datasets instead, the performance of the algorithm on the full dataset seems higher in the validation set of the model trained on the complete data index, which confirms that the more informed and curated dataset outperforms the less curated one and still let us ask what it can be done to use this more informed model to guide or infer on low dimension model or data setting as the vast majority of CS data repository before heavy data pre-processing. The results of the 2023 solution show the benefits and the downfall of DL; while the model seems to perform remarkably in the original context, it fails to generalise to another task of the same instance dataset. Overall, DL models can accurately analyse a citizen science repository, but their performance is too connected to the task since no model satisfies the requirements of all three challenges. The classification tactic used in the 2023 dataset is worth studying since it reveals a connection between mean traits and plant species; while the model does not perform well on the PT2024 settings, the winner of this competition further investigates this connection with a multi-head model that mixes regression and classification[2]. The low performance of transfer learning from the high accuracy model [13] indicates a significant task shift; features learned over classification seem incompatible with features for plant disease, at least in the PT2024 setting since both adapted classification solution as shown fail to give a good performance. The overall structure of a minimal adaptive ensemble appears to perform in scale; the fully trained network with reference architectures outperforms the weak model and the fine-tuned model; even more promising is the increased accuracy of both the informed model and the minimal informed model; underlying the possibility of feature fusion of different data type for adaptive ensembling more complex architecture. The similar performance of the feature extraction model in a complex structure as FT-Transformer and ANN highlights the well-known fact that DL models struggle to process tabular input data. The exciting part is that the catboost algorithm has similar accuracy to the informed FT ensemble as stated in the FT-transformer paper [21].

V. Conclusion

This research investigates potential weaknesses in citizen science (CS) datasets while exploring the feasibility of domain adaptation within similar domains (plant images) for tasks like regression and classification. Interestingly, the CS collection methods for the three datasets resulted in remarkably consistent outlier percentages, data distributions, and image training set characteristics; this study lays the groundwork for subsequent investigations.

However, domain adaptation appeared unable to learn the new problem even within the same plant image domain. In inductive transfer learning [23], the source task should influence the target task. The low accuracy suggests that plant diseases might not directly relate to plant traits, requiring further investigation of this relationship.

The most intriguing finding might be the possibility and coherence of a minimally concatenated dataset. With ongoing research on mixture-of-experts [24] concatenation, a model trained on various assembled datasets with diverse input dimensions is a promising avenue for exploration, especially in settings such as federated learning.

References

[1] J. Silvertown, "A new dawn for citizen science," *Trends in ecology & evolution*, 2009. doi: 10.1016/j.tree.2009.03.017

[2] R. Bonney, C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk, "Citizen science: a developing tool for expanding science knowledge and scientific literacy," *BioScience*, 2009. doi: 10.1525/bio.2009.59.11.9

[2] https://github.com/dysdsyd/PlantTraits2024

[3] C. L. of Ornithology, "eBird," last retrieved July 24, 2024. [Online]. Available: https://science.ebird.org/en

[4] K. O'Donnell, "iNaturalist," last retrieved July 24, 2024. [Online]. Available: https://www.inaturalist.org/

[5] K. S. Chris Lintott, "Zooniverse," last retrieved July 24, 2024. [Online]. Available: https://www.zooniverse.org/

[6] C. Schiller, S. Schmidtlein, C. Boonman, A. Moreno-Martínez, and T. Kattenborn, "Deep learning and citizen science enable automated plant trait predictions from photographs," *Scientific Reports*, 2021. doi: 10.1038/s41598-021-95616-0

[7] M. J. Feldman, L. Imbeau, P. Marchand, M. J. Mazerolle, M. Darveau, and N. J. Fenton, "Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review," *PloS one*, 2021. doi: 10.1371/journal.pone.0234587

[8] S. Wolf, M. D. Mahecha, F. M. Sabatini, C. Wirth, H. Bruelheide, J. Kattge, Á. Moreno Martínez, K. Mora, and T. Kattenborn, "Citizen science plant observations encode global trait patterns," *Nature Ecology & Evolution*, 2022. doi: 10.1038/s41559-022-01904-x

[9] O. Atkin, J. Kattge, S. Diaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bonisch, E. Garnier, M. Westoby, P. B. Reich *et al.*, "Try-a global database of plant traits," *Global Change Biology*, 2011. doi: 10.1111/j.1365-2486.2011.02451.x

[10] WorldClim, "WorldClim," last retrieved July 24, 2024. [Online]. Available: https://www.worldclim.org/

[11] T. Kattenborn, "Planttraits2023," 2023, competition. [Online]. Available: https://kaggle.com/competitions/planttraits2023

[12] A. Awsaf, H.-J. Sharma, M. Görner, and T. Kattenborn, "Planttraits2024 - fgvc11," 2024, competition. [Online]. Available: https://kaggle.com/competitions/planttraits2024

[13] A. Bruno, D. Moroni, and M. Martinelli, "Efficient deep learning approach for olive disease classification," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*, 2023. doi: 10.15439/2023F4794

[14] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "Convnext v2: Co-designing and scaling convnets with masked autoencoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. doi: 10.1109/CVPR52729.2023.01548

[15] C. Schiller, "CNN Models, metadata and global trait distribution maps," dataset Repository. [Online]. Available: https://figshare.com/articles/dataset/CNN_Models_metadata_and_global_trait_distribution_maps/13312040

[16] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939785

[17] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, 2018. doi: https://doi.org/10.48550/arXiv.1706.09516

[18] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. doi: 10.1109/ICCV.2019.00612

[19] A. Bruno, D. Moroni, R. Dainelli, L. Rocchi, S. Morelli, E. Ferrari, P. Toscano, and M. Martinelli, "Improving plant disease classification by adaptive minimal ensembling," *Frontiers in Artificial Intelligence*, 2022. doi: 10.3389/frai.2022.868926

[20] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," *Frontiers in plant science*, 2016. doi: 10.3389/fpls.2016.01419

[21] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," *Advances in Neural Information Processing Systems*, 2021. doi: 10.48550/arXiv.2106.11959

[22] P. Soroye, T. Newbold, and J. Kerr, "Climate change contributes to widespread declines among bumble bees across continents," *Science*, 2020. doi: 10.1126/science.aax8591

[23] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, 2019. doi: 10.48550/arXiv.1911.02685

[24] Z.-A. Huang, Y. Hu, R. Liu, X. Xue, Z. Zhu, L. Song, and K. C. Tan, "Federated multi-task learning for joint diagnosis of multiple mental disorders on mri scans," *IEEE Transactions on Biomedical Engineering*, 2022. doi: 10.1109/TBME.2022.3210940