Cooperative learning of Pl@ntNet's Artificial Intelligence algorithm using label aggregation

Tanguy Lefort [1] & Antoine Affouard [2] & Pierre Bonnet [3] &
Benjamin Charlier [4] & Alexis Joly [5] & Joseph Salmon [5]

[1] *Univ. Montpellier, CNRS, IMAG, Inria, LIRMM, France tanguy.lefort@umontpellier.fr*
[2] *IRD, AMAP, Montpellier, France antoine.affouard@cirad.fr*
[3] *CIRAD, AMAP, Montpellier, France*
[4] *Univ. Montpellier, CNRS, IMAG, France benjamin.charlier@umontpellier.fr*
[5] *Inria, LIRMM, France, alexis.joly@inria.fr*
[6] *Univ. Montpellier, CNRS, IMAG, IUF, France joseph.salmon@umontpellier.fr*

Résumé. Le système Pl@ntNet collecte des données à l'échelle mondiale en permettant aux utilisateurs de télécharger et d'annoter des observations de plantes. Les étiquettes ainsi obtenues bruitées en raison des compétences diverses des utilisateurs. L'obtention d'un consensus est cruciale pour entraîner des modèles d'apprentissage, mais l'échelle des données collectées rend les stratégies traditionnelles d'agrégation des étiquettes difficiles à mettre en œuvre. En outre, comme de nombreuses espèces sont rarement observées, l'expertise des utilisateurs ne peut pas être évaluée comme un accord entre utilisateurs : sinon, les experts en botanique auraient un poids plus faible dans l'étape d'apprentissage que l'utilisateur moyen de part leur participation moindre mais plus ciblée. La stratégie d'agrégation d'étiquettes que nous proposons vise à entraîner de manière coopérative des modèles d'apprentissage automatique pour l'identification des plantes. Cette stratégie estime l'expertise des utilisateurs sous la forme d'un score de confiance par travailleur, basé sur leur capacité à identifier des espèces végétales à partir de données collectées par la foule. Le score de confiance est estimé récursivement à partir des espèces correctement identifiées compte tenu des étiquettes estimées actuelles. Ce score interprétable exploite les connaissances des experts en botanique et l'hétérogénéité des utilisateurs. Nous évaluons notre stratégie sur un large sous-ensemble de la base de données Pl@ntNet axée sur la flore européenne, comprenant plus de 6 000 000 d'observations et 800 000 utilisateurs. Nous démontrons que l'estimation des compétences des utilisateurs basée sur la diversité de leur expertise améliore la performance de l'étiquetage.

Mots-clés. Apprentissage coopératif, aggrégation d'étiquettes, annotation de données, écologie

Abstract. The Pl@ntNet system enables global data collection by allowing users to upload and annotate plant observations, leading to noisy labels due to diverse user skills. Achieving consensus is crucial for training, but the vast scale of collected data makes traditional label aggregation strategies challenging. Additionally, as many species are rarely observed, user expertise can not be evaluated as an inter-user agreement: otherwise, botanical experts would have a lower weight in the training step than the average user as they have fewer but precise participation. Our proposed label aggregation strategy aims to cooperatively train plant identification models. This strategy estimates user expertise as a trust

score per worker based on their ability to identify plant species from crowdsourced data. The trust score is recursively estimated from correctly identified species given the current estimated labels. This interpretable score exploits botanical experts' knowledge and the heterogeneity of users. We evaluate our strategy on a large subset of the Pl@ntNet database focused on European flora, comprising over 6 000 000 observations and 800 000 users. We demonstrate that estimating users' skills based on the diversity of their expertise enhances labeling performance.

Keywords. Crowdsourcing, label aggregation, data annotation, ecology

1 Introduction

Computer vision models are a great aid in plant species recognition in the field [20, 1]. However, to train them we need large annotated datasets. These datasets are often created thanks to citizen science approaches, collecting both reliable and useful information [2]. Among existing plant recognition applications, the Pl@ntNet system enables global data collection by allowing users to upload and annotate plant observations.

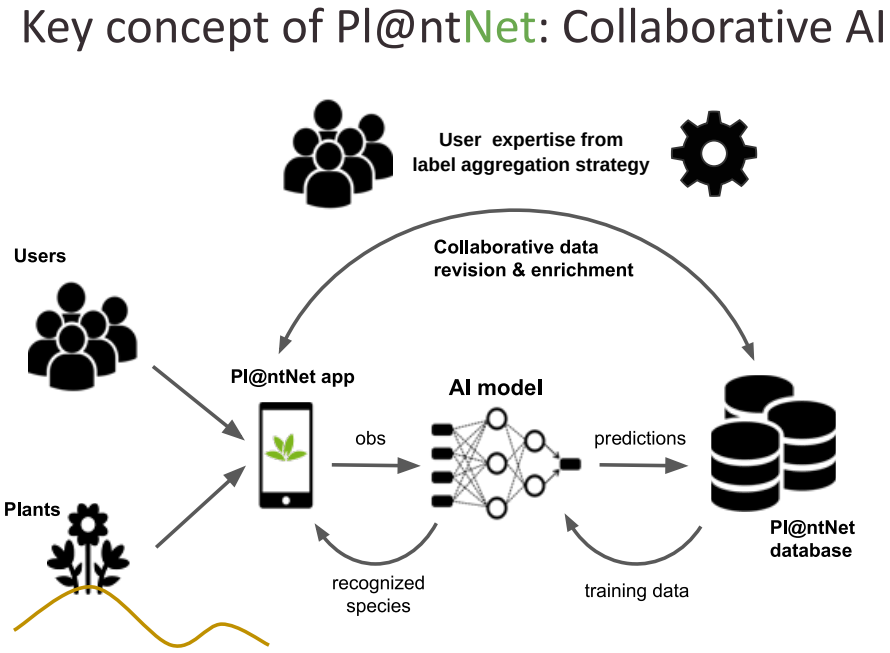


Figure 1: Pl@ntNet system for plant species recognition. Users take their plant observations in the Pl@ntNet application. A prediction is output by the neural network model. Users can validate the prediction or propose another species. The whole votes collection is used to evaluate user expertise (see Algorithm 1) and actively revise observations identifications.

At the time of writing, this participatory approach has resulted in the collection of over 20 million observations (image or group of images of a same plant), belonging to almost

46 000 species, by more than 6 million observers worldwide. The collaborative process of Pl@ntNet is described in Figure 1. The model interacts with the human decision by proposing possible species given an observation. For each returned species, using a similarity search, the Pl@ntNet system also shows similar pictures from the database. This lets users to visually check that their observation is likely to belong to a predicted species given the most similar observations. Such a visual control can be of help when flowering is not yet complete to compare two plants at similar growth stages. Plant species identification is a task that requires skills to recognize morphological traits (shapes, measurements, environments and specific characteristics). A large number of users with diverse skills have participated in gathering plant observations and helped improve the training dataset of our computer vision model. Their participation is based on votes that they can cast on others observations, or by the initial species determination of their observation. The quality of each vote is then processed by the algorithm presented in Section 2.2.

Other citizen science projects such as iNaturalist [19] or eBird [18] use a similar approach to collect data. However, each platform has its own label aggregation strategy. The iNaturalist project, with more than 2.5 million users, records the votes at different taxonomic levels. The resulting label is the aggregation of at least two votes on a species-level identification (or coarser or finer taxonomic level) and the taxon needs at least two-thirds of identifier agreements – in particular, all users have the same weight in the decision-making. Over time, a taxon can be further refined by the community or revoked. eBird handles taxon quality control by using a checklist in each region for observers. Quality verifications on the checklist are performed and, combined with user knowledge – the number of species and checklist submitted, number of flagged observations, further discussions with local experts – the observation taxon is accepted. The eBird project also showed that monitoring species accumulation from observers can help to sort their skills [10]. While they consider the species accumulation by hours spent on each collected observation, we propose a strategy that takes into account the entire history of observations of the observer.

In this article, we present the Pl@ntNet label aggregation strategy. Using a large-scale dataset of more than 6 million observations and 800 thousand users, we show that our strategy can improve the quality of the collected data, without removing every observation that was only labeled by single users. This work is ongoing and the dataset will be released with codes.

2 Methods

2.1 Dataset and notation

To compare the different label aggregation strategies on large-scale datasets, we consider a subset of the Pl@ntNet database focused on Southwestern European flora observations – Baleares, Corsica, France, Portugal, Sardegna and Spain – from 2017 to October 2023. In total, 9 005 108 votes are cast by $n_{\text{user}} = 823\,251$ users on 6 699 593 observations after cleaning steps. Those cleaning steps include filtering out identification votes with proposed plant species not available in the World Checklist of Vascular Plants (WCVP) [6]. Thanks

to Kew's Royal Botanical Garden, we adopted the Plants of the World Online [15] system with the *k-southwestern-europe*. Within this taxonomic checklist, we removed synonyms. However, there are plant species listed in *k-southwestern-europe* POWO that are not in the WCVP checklist. As there is a possible taxon ambiguity in this case – multiple species possible for a given synonym depending on the referential – we leave the proposed label untouched.

Notation In the following, denote $K = 11\,425$ the number of species within the dataset. We index the observations by $i \in [n_\bullet] = \{1, \dots, n_\bullet\}$ where $\mathcal{D}_\bullet$ is the considered dataset composed of $n_\bullet$ observations and their associated votes. For example, the full south-western european flora dataset from Pl@ntNet of $6\,699\,593$ observations is denoted $\mathcal{D}_{\text{SWE}}$. Other subsets are presented in Section 2.3. We write $\mathcal{U}$ the set of users. Each user u has a unique identifier used as an index, and we denote $\mathcal{U}_i$ the set of users that have voted on observation i – *i.e.* $\mathcal{U} = \cup_{i \in [n_{\text{SWE}}]} \mathcal{U}_i$. The vote of user u on observation i is denoted $y_i^u \in [K]$. Each observation i is created by an author u stored in Author(i).

2.2 Proposed label aggregation strategy

Pl@ntNet label aggregation strategy relies on estimating the number of correctly identified species for each user. Similar to other strategies, we rely on an EM based iterative procedure [5] to estimate consecutively the users' skills and each observation's species. As the collected data is used to train the model, the label aggregation strategy also generates a trust indicator on the observation. This quality indicator reveals if the observation is valid or not. The AI model is then only trained on valid observations. This operation is done monthly to keep the system up-to-date with the latest observations. The more users vote on observations, the more valid observations are identified and the better the model. Notice that proposing a species as author of the observation weighs ten times more than voting by click in Algorithm 1. Indeed, being on the field leads to more information on the environment and a better determination of the species. Finally, note that species are unequivocally identified as author's (n_u^{author} in Algorithm 1) or as votes on other's observations (n_u^{vote}) in the aggregation strategy. The final number of species identified by users is the aggregation of these two terms: $n_u = \text{Round}\left(n_u^{\text{author}} + \frac{1}{10} n_u^{\text{vote}}\right)$.

From Algorithm 1, we see that a user becomes **self-validating** (*i.e.* trusted enough so that their label checks observations as valid identifications) when their weight w_u is greater than the level θ_{conf}. In practice, this means that an experienced user who has collected enough weight can validate any observation without any other user's vote. Note that this identification can later be invalidated by other users with enough weight thanks to the accuracy threshold θ_{conf}. Moreover, the weight function f shown in Figure 2 is a non-decreasing function that maps the number of identified species n_u to a trust score in the form of:

$$w_u = f(n_u) = n_u^\alpha - n_u^\beta + \gamma \ , \tag{1}$$

where $\alpha, \beta \in \mathbb{R}_+^\star$ are hyperparameters that were calibrated internally to fit prior knowledge and $\gamma > 0$ is the constant representing the initial weight of each user. In practice, we use

Algorithm 1 Pl@ntNet label aggregation strategy

Input: Votes as $(u, y_i^u)_{i\in[n_{\text{SWE}}],u\in[n_{\text{user}}]}$ for each observation i and user u answering the voted species y_i^u, accuracy threshold θ_{acc}, confidence threshold θ_{conf}, weight function f, initial weight $\gamma > 0$

Output: Estimated labels $\hat{y}_i$ for each observation i

Initialize $\hat{y}_i = \text{MV}\left(\{y_i^u\}_u\right)$ for each observation $i \in [n_{\text{SWE}}]$

Initialize user weights as $w_u = \gamma$ for each user $u \in [n_{\text{user}}]$

while not converged **do**

  for each observation $i \in [n_{\text{SWE}}]$ **do**

    Compute label confidence: $\text{conf}_i(\hat{y}_i) = \sum_{u\in\mathcal{U}_i} w_u \mathbb{1}(y_i^u = \hat{y}_i)$

    Compute label accuracy: $\text{acc}_i(\hat{y}_i) = \text{conf}_i(\hat{y}_i) / \sum_{k\in[K]} \text{conf}_i(k)$

    Compute validity indicator: $s_i = \mathbb{1}(\text{acc}_i(\hat{y}_i) \geq \theta_{\text{acc}} \text{ and } \text{conf}_i(\hat{y}_i) \geq \theta_{\text{conf}})$

  end for

  for each user $u \in [n_{\text{user}}]$ **do**

    Compute the number of valid identified species for authoring observations:

$$n_u^{\text{author}} = |\{y_i^u \in [K] \,|\, y_i^u = \hat{y}_i, s_i = 1, \text{Author}(i) = u\}|$$

    Compute the number of identified species by voting on other's observations:

$$n_u^{\text{vote}} = |\{y_i^u \in [K] \,|\, y_i^u = \hat{y}_i, \text{Author}(i) \neq u\}|$$

    Compute the rounding number of identified species per user:

$$n_u = \text{Round}\left(n_u^{\text{author}} + \frac{1}{10} n_u^{\text{vote}}\right)$$

    Transform number of estimated species per user into trust score: $w_u = f(n_u)$

  end for

  Update estimated labels with a weighted majority vote

$$\forall i \in [n_{\text{SWE}}],\ \hat{y}_i = \underset{k\in[K]}{\arg\max} \sum_{u\in\mathcal{U}_i} w_u \mathbb{1}(y_i^u = k)$$

end while

$\alpha = 0.5$, $\beta = 0.2$ and $\gamma = \log(2.1) \simeq 0.74$ in the weight function. As for the two thresholds that control the level of uncertainty accepted for a given label, they are set to $\theta_{\text{conf}} = 2$ to control the total weight on an observation and $\theta_{\text{acc}} = 0.7$ to control the agreement between users given their expertise.

2.3 Evaluation against other aggregation strategies

Existing aggregation strategies Plant species label aggregation is a challenging task due to the large number of species K. Hence, many classical strategies in the label aggregation

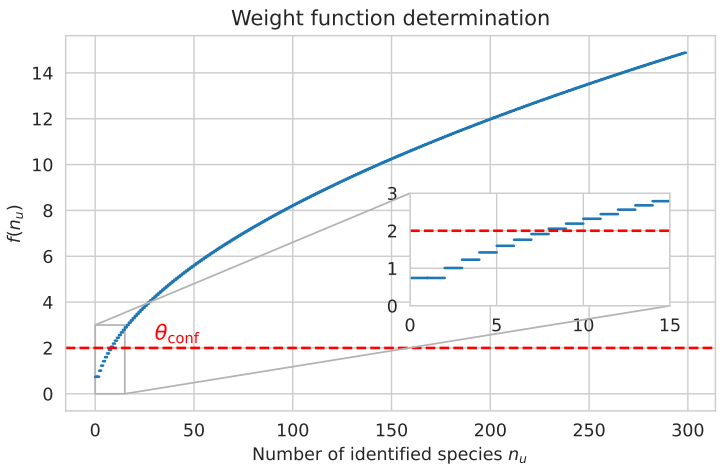


Figure 2: Weight function (Equation (1)) used to map the number of identified species to a trust score in the Pl@ntNet label aggregation strategy. The user confidence threshold $\theta_{\text{conf}} = 2$ requires a user to have identified at least $n_u = 8$ species to become **self-validating**. A new user starts with a weight of $f(0) = f(1) = \gamma \simeq 0.74$.

literature such as Dawid and Skene's [4] and other variations [14, 17] are not applicable as they require estimating a K^2 matrix for each worker. This would result, in the considered dataset $\mathcal{D}_{\text{SWE}}$ as $11\,425^2 \times 823\,251 \approx 10^{14}$ parameters to estimate. Similar issues occur for other label aggregation strategies [21, 8, 13]. We do not consider deep-learning based crowdsourcing strategies as Rodrigues and Pereira [16] and Chu, Ma, and Wang [3] or Lefort et al. [11] as they train a model from crowdsourced labels but do not output aggregated labels on the training set. In the Pl@ntNet application, we need to propose one or multiple species for each observation to users. To overcome these issues, we consider the following label aggregation strategies that can scale to large K and number of users:

- **Majority Vote (MV)[9]**: Certainly the most common aggregation strategy, the majority vote selects the most answered label. In the case of equalities, a random draw is performed – creating sometimes some variability in the labeling process. More formally, given an observation i:

$$\text{MV}(i, \{y_i^u\}_u) = \arg\max_{k \in [K]} \sum_{u \in \mathcal{U}_i} \mathbb{1}(y_i^u = k) \ .$$

- **Worker agreement with aggregate (WAWA)** [12]: Also known as the inter-rater agreement, this strategy weights each user by how much they agree with the MV labels on the images they annotated. More formally, given an observation i:

$$\text{WAWA}(i, \mathcal{D}_{\text{SWE}}) = \arg\max_{k \in [K]} \sum_{u \in \mathcal{U}_i} w_u \mathbb{1}(y_i^u = k)$$

$$\text{with } w_u = \frac{1}{|\{y_{i'}^u\}_{i'}|} \sum_{i'=1}^{n_{\text{SWE}}} \mathbb{1}\left(y_{i'}^u = \text{MV}(\{y_{i'}^u\}_u)\right) \ .$$

- **iNaturalist** [19]: The iNaturalist platform generates a label for observations with at least two votes. The estimated label represents the one with at least two-thirds of the majority in agreement. Every user has the same weight in the aggregation. More formally:

$$\text{iNaturalist}(i, \{y_i^u\}_u) = \begin{cases} \text{MV}(i, \{y_i^u\}_u) & \text{if } s_i = 1 \\ \text{undefined} & \text{otherwise} \end{cases}$$
$$\text{with } s_i = \mathbb{1}\left(\max_{k \in [K]} \frac{1}{|\mathcal{U}_i|} \sum_{u \in \mathcal{U}_i} \mathbb{1}(y_i^u = k) \geq \frac{2}{3} \right) \ .$$

As there is no observation filter for the MV and WAWA, we consider that for all observation i, $s_i = 1$ for these two strategies. Experiments were completed using the `peerannot` library (`https://github.com/peerannot/peerannot`).

Creation of an evaluation set in a crowdsourcing setting To evaluate the performance of a label aggregation strategy, it is necessary to know the ground truth. However, in the context of crowdsourced data, there is no known truth for the observations. The sheer volume of data makes it impossible to ask botanical experts to create such ground truth for the whole database.

Instead of asking such experts to label a subset of the data, we identified botanical experts in our users database. From within the Pl@ntNet team, we referenced well-known botanists to start a list of expert users. To these we have added TelaBotanica [7] users with registered confirmed botanical experience from their directory and that are also Pl@ntNet users that participated to the South-Western Europe flora subset. Among the users, 98 are identified as botanical experts by the Pl@ntNet team and Telabotanica platform. The answers of these experts are considered as ground truth labels and used to evaluate strategies performance. Despite our selection process of supposedly "indisputable" experts, a few observations in the test set denoted $\mathcal{D}_{\text{expert}}$ still end up with contradictory labels (4 observations in total). As they represent a very small percentage, we simply removed them from $\mathcal{D}_{\text{expert}}$.

Our evaluation set $\mathcal{D}_{\text{expert}}$ is finally composed of 26 811 observations. Of these evaluation data, 17 125 received more than two identifications and are stored in $\mathcal{D}_{\text{multiple votes}}$; 1 263 have more than two votes with at least one disagreement between users are stored in $\mathcal{D}_{\text{disagreement}}$. Figure 3 shows the distribution of observations from $\mathcal{D}_{\text{SWE}}$ to the finer and more ambiguous $\mathcal{D}_{\text{disagreement}}$.

Unfortunately, the demand for multiple labels on observations is not being met, despite the large number of users. Indeed, 310 564 users were single time voters (meaning they interacted with the system only once).

Evaluation metric To evaluate the label aggregation strategies, we use the following accuracy metrics computed on valid observations ($s_i = 1$):

$$\text{Acc}(\hat{y}, y; \mathcal{D}_\bullet) = \frac{1}{n_\bullet} \sum_{i=1}^{n_\bullet} \mathbb{1}(\hat{y}_i = y_i)\mathbb{1}(s_i = 1) \ ,$$

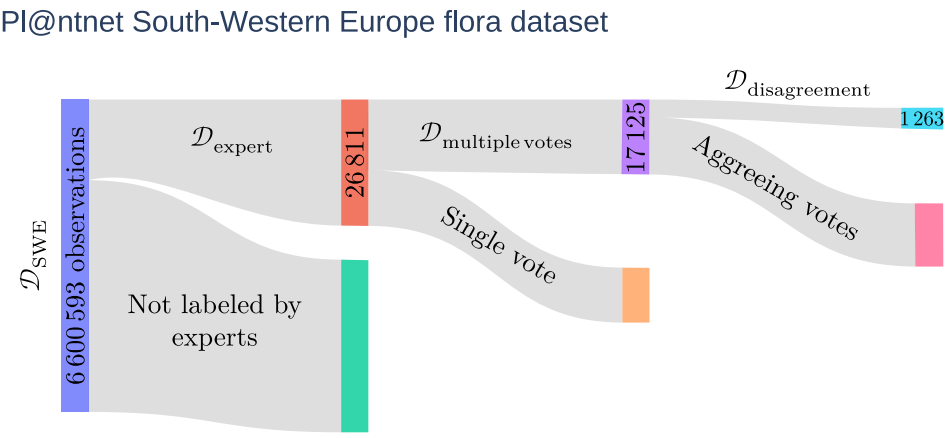


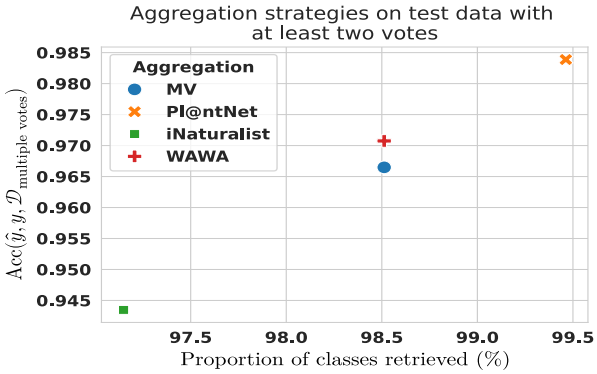
Figure 3: Log-scales distribution of the observations in the South-West European Flora subset from the Pl@ntNet database.

with $\hat{y} = (\hat{y}_i)_i$ the estimated labels on the considered set $\mathcal{D}_\bullet \subset \mathcal{D}_{\text{expert}}$, $y = (y_i)_i$ the associated experts labels, considered as ground truth. When the aggregation strategy indicates the observation as invalid ($s_i = 0$ for Pl@ntNet and iNaturalist), we consider the label as incorrect in the performance measure as an expert was able to decide on a species. Finally, we also consider the proportion of species retrieved by the aggregation strategies. This is important as if a species identified by the experts disappears during the aggregation, the model trained from this aggregated data can no longer predict this species.

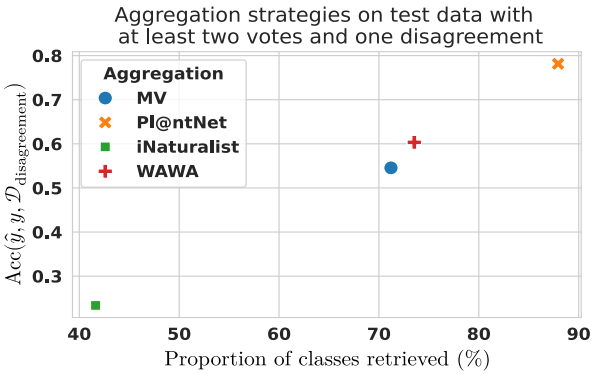
We evaluate the label recovery of each strategy on three subsets visualized in Figure 3: the full test set where experts have voted a species, the subset of observations with at least 2 votes and the subset of observations with at least 2 votes and one disagreement. The latter subset is the most challenging as it contains the observations with the most ambiguity. We selected these subsets to investigate the label aggregation strategies' performance depending on the ambiguity level.

3 Results

Accuracy of the aggregation strategies We evaluate the accuracy of the strategies on the set of tasks labeled by our experts. Figure 4 shows how many predicted labels match our experts answers on $\mathcal{D}_{\text{multiple votes}}$ and $\mathcal{D}_{\text{disagreement}}$. More importantly, we compare this quantity with the volume of class retrieved by the aggregation strategy. We observe that the data filtering from the iNaturalist strategy impacts its performance. On $\mathcal{D}_{\text{expert}}$, MV reaches 97% of accuracy, WAWA 98%, iNaturalist 60% and Pl@ntNet 99%. To differentiate between the best performing strategies, we need to look at more ambiguous observations like those in $\mathcal{D}_{\text{multiple votes}}$ and $\mathcal{D}_{\text{disagreement}}$. In high ambiguous frameworks, the WAWA strategy outperforms the MV one. However, overall the Pl@ntNet aggregation is more often in adequation with the experts and retrieves almost 90% of plant species identified by experts in high ambiguous datasets against 73% for WAWA, 71% for MV and only 41% for iNaturalist.



(a) Accuracy on $\mathcal{D}_{\text{multiple votes}}$ against volume of species recovered



(b) Accuracy on $\mathcal{D}_{\text{disagreement}}$ against volume of species recovered

Figure 4: Accuracy of the aggregation strategies against the volume of class retrieved on subsets with at least two votes – either agreeing (A) or with at least one disagreeing vote (B). The Pl@ntNet aggregation is more accurate especially in a highly ambiguous setting (B). The iNaturalist data filter highly impacts how many classes are kept in the dataset and the overall accuracy in both settings. WAWA and MV perform similarly with a benefit for WAWA when skill evaluation is needed.

4 Conclusion

We demonstrated that collaborative identification of plant species can effectively be used to obtain expert levels labels. Using a large subset of millions of observations and thousands of users from the Pl@ntNet organization, we investigate a label aggregation strategy that weighs user answers based on their estimated number of species correctly identified without using prior expert knowledge. Many strategies used previously either do not scale to the magnitude of the current databases – either Pl@ntNet, iNaturalist or eBird – or are outperformed by our aggregation. Our strategy weighs users based on the number of correctly identified species. This weight is interpretable and shows the diversity of the user's skillset. It can be directly applied on other crowdsourced frameworks with a high number of classes like iNaturalist's.

References

[1] M. L. Borowiec et al. "Deep learning as a tool for ecology and evolution". In: *Methods in Ecology and Evolution* 13.8 (2022), pp. 1640–1660.

[2] E. D. Brown and B. K. Williams. "The potential for citizen science to produce reliable and useful information in ecology". In: *Conservation Biology* 33.3 (2019), pp. 561–569.

[3] Z. Chu, J. Ma, and H. Wang. "Learning from Crowds by Modeling Common Confusions." In: *AAAI*. 2021, pp. 5832–5840.

[4] A. P. Dawid and A. M. Skene. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". In: *J. R. Stat. Soc. Ser. C. Appl. Stat.* 28.1 (1979), pp. 20–28.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 39.1 (1977), pp. 1–22.

[6] R. Govaerts. *The World Checklist of Vascular Plants (WCVP)*. Checklist dataset. Accessed via GBIF.org on 2024-01-30. 2023. DOI: 10.15468/6h8ucr.

[7] L. Heaton, F. Millerand, and S. Proulx. "« Tela Botanica » : une fertilisation croisée des amateurs et des experts". In: *Hermès, La Revue* 57.2 (2010), pp. 61–68.

[8] D. Hovy et al. "Learning Whom to Trust with MACE". In: *Proceedings of NAACL-HLT 2013*. 2013.

[9] G. James. "Majority vote classifiers: theory and applications". PhD thesis. Stanford University, 1998.

[10] S. Kelling et al. "Can Observation Skills of Citizen Scientists Be Estimated Using Species Accumulation Curves?" In: *PLOS ONE* 10.10 (Oct. 2015), pp. 1–20.

[11] T. Lefort et al. "Identify ambiguous tasks combining crowdsourced labels by weighting Areas Under the Margin". In: *arXiv preprint arXiv:2209.15380* (2022).

[12] A. Limited. *Calculating Worker Agreement with Aggregate (Wawa)*. 2021. URL: https://success.appen.com/hc/en-us/articles/202703205-Calculating-Worker-Agreement-with-Aggregate-Wawa-.

[13] Q. Ma and A. Olshevsky. "Adversarial crowdsourcing through robust rank-one matrix completion". In: *NeurIPS*. Vol. 33. 2020, pp. 21841–21852.

[14] R. J. Passonneau and B. Carpenter. "The Benefits of a Model of Annotation". In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 311–326.

[15] POWO. *Plants of the World Online*. Published on the Internet. Facilitated by the Royal Botanic Gardens, Kew. 2024. URL: http://www.plantsoftheworldonline.org/.

[16] F. Rodrigues and F. Pereira. "Deep learning from crowds". In: *AAAI*. Vol. 32. 2018.

[17] V. B. Sinha, S. Rao, and V. N. Balasubramanian. "Fast Dawid-Skene: A fast vote aggregation scheme for sentiment classification". In: *arXiv preprint arXiv:1803.02781* (2018).

[18] B. Sullivan et al. "eBird: A citizen-based bird observation network in the biological sciences". In: *Biological conservation* 142.10 (2009), pp. 2282–2292.

[19] G. Van Horn et al. "The inaturalist species classification and detection dataset". In: *CVPR*. 2018, pp. 8769–8778.

[20] M. Vidal et al. "Perspectives on individual animal identification from biology and computer vision". In: *Integrative and comparative biology* 61.3 (2021), pp. 900–916.

[21] J. Whitehill et al. "Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise". In: *NeurIPS*. Vol. 22. 2009. (Visited on 2021).