A First Approach for Handling Uncertainty in Citizen Science

Manuel Jiménez[1], Isaac Triguero, *Member, IEEE*[1] and Robert John, *Senior Member, IEEE*[1,2]

[1]Automated Scheduling, Optimisation and Planning (ASAP)
[2]Laboratory for Uncertainty in Data and Decision Making (LUCID)
School of Computer Science, University of Nottingham, UK
Email: {manuel.jimenezmorales, isaac.triguero, robert.john}@nottingham.ac.uk

***Abstract*—Citizen Science is coming to the forefront of scientific research as a valuable method for large-scale processing of data. New technologies in fields such as astronomy or bio-sciences generate tons of data, for which a thorough expert analysis is no longer feasible. In contrast, communities of volunteers coordinated by the Internet are showing a great potential in completing such analysis in a reasonable time. However, this approach brings uncertainty and the spread of biases within the data, since amateur participants are usually non-experts on the subject and count with variable skills and expertise. This means lack of accuracy in results coming from Citizen Science projects.**

This work presents a novel approach to handle uncertainty in Citizen Science. We focus on leveraging this uncertainty in the data pursuing a refinement of results. We distinguish between two types of uncertainty: a first one due to the lack of consensus between amateurs, and another one quantified by amateurs themselves during the course of the project. We test our method using the Galaxy Zoo, a project which aims for the labelling of a huge dataset of galaxy images. Considering available expert classifications to validate our experiments, the proposed method is able to improve current accuracy and classify a greater number of images.

I. Introduction

In these days, connectivity among people all around the globe is boosting the emergence of a great potential for tackling complex problems. Whereas in some cases this complexity restricts the problem analysis to a set of experts, for many others a novel approach is leveraging tiny efforts from a huge amount of amateur people. We refer to Citizen Science as the development of scientific research assisted by volunteers from the general public [1]. Within a crowdsourcing context, this practice is flourishing as a promising tool for problems that entail the processing of huge volumes of data in the form of high time-consuming tasks such as labelling of images, gathering of environmental records, or transcription of handwritten texts. In particular, the nascent discipline of astroinformatics [2] is benefitting from the analysis of astronomical data in multiple projects. However, Citizen Science outreach goes beyond that particular application, covering problems in many other fields such as ecology or bioinformatics [3].

Citizen Science has also been attracting the attention of data science research along the past decade. From a practical point of view, amateurs engaged in a Citizen Science project can be regarded as a massive set of imperfect classifiers. Several studies have reported the use of amateur-labelled data obtained after the project closure, using an off-line approach. Most of them have focused on the performance of Machine Learning (ML) algorithms using this data [4], [5]. Recently, ML applications have also been implemented for the optimisation of Citizen Science on-line platforms [6]. In this work, however, we are focusing on an off-line approach, that is, the mining of the Citizen Science data in order to improve the accuracy reached by amateurs on themselves.

Despite all of this, Citizen Science arouses scepticism within the scientific community [7]. Although it enables data analysis at scales not possible to accomplish by professional researchers on their own, the practice is not universally accepted as a valid method for scientific research [8]. The main concern is the quality of the results, commonly questioned because of the prevalence of biases and lack of accuracy [9]. Volunteers do not generally count with any background in sciences or research, and depending on the difficulty of the task and their expertise, amateur classifications broadly vary in their level of confidence. Consequently, Citizen Science data holds an intrinsic uncertainty, which can be alleviated with appropriate analyses and the aid of expert knowledge about the problem addressed in each case. In this work, we begin to explore the use of fuzzy logic along with the help of expert knowledge in order to embrace the uncertainty within the Citizen Science data.

To date, fuzzy logic has provided good results for the encapsulation of expert knowledge in several domains. Different works address this issue in the frame of Multi-criteria Decision Making and Multi-expert Decision Making, when there is available a set of choices and a variable range of expert judgements [10], [11]. Nonetheless, to the best of our knowledge, these approaches have not been applied to amateur knowledge in Citizen Science projects. Unlike in Multi-expert Decision Making, Citizen Science involves vast amounts of data annotated by a great number of non-experts on the subject.

In this paper, we propose an integrated use of data labelled by amateurs in the course of a Citizen Science project. We aim to provide a framework to handle uncertainty in the data resulting from Citizen Science projects that maximises the potential utility of Citizen Science outcomes for research. In our experiments, we will focus on the first edition of the

Galaxy Zoo (GZ1) project [12], as the very first successful implementation of Citizen Science using the Internet. GZ1 came up with morphological classifications for nearly a million galaxies with the support of more than 200,000 amateur participants. The proposed approach presents an innovative refinement of the data by leveraging the inherent uncertainty when multiple independent non-expert judgements come into play. As a result, it is able to provide better classifications for a greater number of galaxies, improving the results previously obtained by professional astronomers. This is tested using available expert classifications on a subset of the same images.

The paper is structured as follows. In Section II, we introduce the background needed for the understanding of the paper. In Section III we present our approach for the refinement of Citizen Science data. Section IV presents the whole experimental setting along with the discussion of results. Finally, in Section V we draw some conclusions and outline possible directions for future work.

II. Background

In this section, we extend the two main topics addressed in the paper. First, we explain in more detail several aspects involved in the development of Citizen Science projects and review current trends in the specialised literature (Subsection II-A). After this, we examine fuzzy logic as a promising resource for research in the improvement of Citizen Science data (Subsection II-B).

A. Citizen Science: A quick overview

Citizen Science is not a new practice. It is rooted more than a century ago, when amateurs started making small contributions to the study of meteorology and ornithology with their records. However, global communications have broadened the different ways volunteers can aid today's research, and Citizen Science is being re-discovered by the scientific community. Thousands of projects are engaging tons of individuals through the Internet in collecting and/or analysing scientific data, with support from multiple institutions from research and academia. Among them, the Zooniverse[1] project stands out as one of the main platforms for the development and management of Citizen Science projects. Currently, Zooniverse has more than a million participants all over the world, engaged in more than 60 projects in topics such as space sciences, ecology, medicine and the humanities [13]. All these efforts have led to the publication of more than 250 scholarly articles[2], validating the help of Citizen Science as meaningful resource for research.

When a project is released, volunteers are invited to complete a particular task, carrying out genuine data analysis (Figure 1). This has traditionally involved the classification of large collections of images related to some research problem. Firstly, participants are normally taught to perform the required task via a simple guide or tutorial. After this, they are asked to classify the images, which are randomly displayed. Participants must then choose between a set of categories, by clicking on the web. These categories often include a set of main classes, usually two or three, which are the target of the classification problem. In addition, users are offered other secondary categories. One of these, the *Don't Know* (DK) category, constitutes a first order measure of the uncertainty in the classification of that particular image. This option ensures every time an object is shown it obtains a vote from the participant.

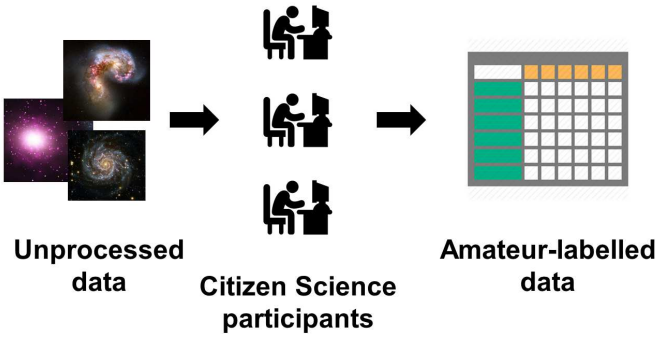

Fig. 1: General view of the Citizen Science workflow.

Several works have tackled Citizen Science as a phenomenon, focusing on the main concerns this approach presents to real research [1], [8], [14]. These studies analyse the potential in the crowd, as a valuable resource for research that should not be neglected by scientific community. This depends on the development of adequate tools for fighting against two main issues brought up by scientists: lack of accuracy and proliferation of biases within Citizen Science results [9]. To this aim, the validation of Citizen Science data is a core aspect that has also been addressed [15].

Data science research applied to Citizen Science has mainly been centred in the use of the resulting data. Off-line approaches offer the possibility of mining the whole datasets, along with additional information regarding participants' performance, expert knowledge on the problem, and other related statistics about the running of the project. Several works have followed the replication of amateur classification skills by using ML algorithms in off-line approaches. It has been shown how ML classifiers are able to achieve comparable results to those obtained by amateurs on themselves, when these algorithms are trained with Citizen Science data [4], [5], [16]. On-line approaches have been recently developed as well, aiming for the best interaction between humans and machines through the running of Citizen Science projects [17]. These approaches pursue to optimise the running of the project. This is targeted using ML systems that enable a progressive train of participants and a synergy between amateur classifications, expert opinions, and automated classifiers [6].

[1] http://www.zooniverse.org

[2] A complete list of publications can be found at http://www.zooniverse.org/publications.

B. Fuzzy logic for decision making: A promising resource for Citizen Science

Fuzzy logic deals with uncertainty in decision problems. This is covered by a set of subjects that face the multiple forms this uncertainty arises, depending on the nature of the problem, and number and quality of decision makers. Eventually, an aggregation method is required in order to combine individual preferences or criteria into a final decision, which is expected to take into consideration all individual contributions. Two of the main decision making paradigms in fuzzy logic are Multi-criteria Decision Making (MCDM) and Multi-expert Decision Making (MEDM) [10], [18].

In MCDM we are concerned with finding the most adequate choice when a set of predetermined alternatives is available. A major issue with this kind of approach, however, is that decision makers exhibit variations in their judgements. This has been addressed via different applications, and considering Type-1 or Type-2 fuzzy sets [11]. In contrast, MEDM problems tackle the encapsulation of expert knowledge, when the set of experts exhibits variation in the decision-making context. This problem has been thoroughly studied in medical domains, when linguistic terms may have different meanings for different experts, and their interpretations may also vary depending on the environmental conditions or over time [10].

Research on these problems represents a valuable resource for the improvement of Citizen Science outcomes. Most of the time, amateur participants face MCDM settings depending on the nature of the classification problem proposed. Additionally, MEDM approaches establish a proper context for the aggregation of available expert knowledge to enhance amateur classifications. Given this, here we consider the potential of these fuzzy logic domains and lay out a preliminary study on the use of uncertainty within amateur-labelled data.

III. A Fuzzy Approach for Citizen Science

In this section we present our approach for handling the uncertainty spread within the data collected in a Citizen Science project. Hence, this is an off-line approach, as we take the whole data obtained once the project has finished collecting votes. We first introduce some basic notation for a clear explanation of the method. After this, we present two types of uncertainty proposed for the refinement of classifications using amateur votes. Finally, we illustrate the approach considering a binary classification problem as example.

Once the count of the amateur votes has finished, votes are usually converted into scores by dividing the votes in each category by the total number of votes received by the object. Hence, being $\mathbf{N} = (n_1, n_2, ..., n_{|N|})$ the vote vector, with $|N|$ the number of categories in the problem, we get the score vector $\mathbf{X} = (x_1, x_2, ..., x_{|N|})$ by computing $x_i = \frac{n_i}{M}$, with $M = \sum n_i$ and $i \in \{1, 2, ..., |N|\}$. Traditionally, the set of score vectors has been employed to compute final classifications applying a threshold over the scores: the category which score is greater or equal than the threshold is assigned to the object. The advantage of this procedure is that we can adjust the confidence in the labels. However, the selection of the threshold is completely arbitrary, and all objects that do not reach the threshold are thus labelled as *uncertain*. This makes the classification ineffective as we require more accurate results: the higher the threshold is, the larger is the set of *uncertain* objects.

The main issue concerning the use of Citizen Science data lies in the pervasive uncertainty when a group of people has provided classifications about the same object. Moreover, there is an additional variability in the total number of votes received, M. Amateur judgements are not expected to agree, and final labels thus depend on how this disparity is tackled. The final aim of the method proposed here is to refine these classifications by leveraging this lack of consensus among Citizen Science participants. Our approach improves these classifications in two ways: on the one hand, accuracy is boosted with respect to the benchmark obtained by experts on the problem. On the other hand, it provides classifications for many objects relegated as *uncertain* so far.

We consider two types of uncertainty:

- In the first instance, we refer to *inherent uncertainty* (IU) as the uncertainty in the classification due to inherent variation in amateurs' judgements. As it is explained above (Section II), each amateur performs a vote (click) according to his/her opinion about the object displayed in the website at the time. Amateurs vote on their own and, consequently, the final outcome consists of a record of the clicks counted in each category. Depending on the spread of this count of votes, we can say that the object being classified holds more or less IU. Ideally, an object with all votes grouped in one category holds the least possible IU. Conversely, an object with the votes equally split across all categories means the highest IU.
- Different than the IU, we refer to *measured uncertainty* (MU) as the uncertainty determined itself by amateurs every time they vote the object by clicking on the DK category. Citizen Science projects that tackle classification problems typically have such category. These votes represent a direct measure of the uncertainty in the classification of the object, and it is independent of the spread in the count of votes measured by IU. As a result, an object with zero DK votes holds the least possible MU. In contrast, an object with all its votes in such category shows the greatest MU possible.

Our model takes as input the whole set of scores as well as the count of votes, that is, $\mathbf{X}$ and $\mathbf{N}$ vectors for each object in the dataset. Through our experiments, we also apply a threshold in order to obtain final classifications. However, we consider both IU and MU to insert modifications into the scores and end up classifying more objects. These modifications are developed in two main stages, as follows.

The first stage aims to remove the noise due to secondary categories, so that the application of a threshold does not consider the minority classes and DK votes. It tackles the IU

by a usual normalisation of the main scores. For instance, in a binary classification problem, let $\mathbf{X} = (x_1, x_2)$ be the score vector containing the two main scores, the normalised score vector $\mathbf{Z} = (z_1, z_2)$ is calculated as shown in Equation 1.

$$z_i = \frac{x_i}{\sum x_i}, \quad \text{for} \quad i \in \{1, 2\} \tag{1}$$

This normalisation assures that $z_1 + z_2 = 1$, amending the vagueness induced by DK votes and other secondary categories that might lower the main scores. This mostly occurs in undefined objects with many DK votes and prone to be annotated as *uncertain*.

The second stage aims to aggregate the information contained in the DK votes. It thus tackles the MU by introducing a shift in the previous normalised score vector $\mathbf{Z}$ (Equation 1). This shift is computed using the DK votes owned by the object, n_{DK}, which is weighted using the mean of DK votes across the whole set of examples, μ_{DK}. These two measures are combined as shown in Equation 2, using two additional parameters, α and β, to be adjusted empirically using the original scores.

$$\epsilon = \frac{\alpha \cdot \mu_{DK}}{\beta + n_{DK}} \tag{2}$$

The shift (ϵ) is aggregated to the normalised scores so that the normalisation is preserved, that is, they remain adding the unit. Considering again a binary classification problem, with $\mathbf{Z} = (z_1, z_2)$ the normalised score vector obtained from Equation 1, the new shifted score vector $\mathbf{W} = (w_1, w_2)$ is calculated as indicated in Equation 3, with ϵ obtained in Equation 2.

$$\begin{cases} w_1 = z_1 + \epsilon \\ w_2 = z_2 - \epsilon \end{cases} \tag{3}$$

The novelty of this approach lies in two aspects. Firstly, the normalisation of the scores (Equation 1) turns fuzzy the uncertainty in the classification due to lack of agreement between amateurs. One object with $\mathbf{X} = (1, 0)$ or $\mathbf{X} = (0, 1)$ exhibits the greatest confidence in belonging to classes represented by x_1 or x_2 scores, respectively. On the contrary, objects with $\mathbf{X} = (0.5, 0.5)$ show the least confidence and greatest IU. Secondly, the consideration of MU allows for an aggregation of information about the uncertainty in the classification held in DK votes, by modifying the original scores (Equations 2 and 3). In this case, one object with $x_{DK} = 1.0$ holds the greatest possible MU, higher as n_{DK} takes larger values. Conversely, objects with $x_{DK} = 0.0$ present zero MU.

IV. Case Study

In this section we present the case study chosen for the testing of our proposal: the first edition of the Galaxy Zoo (GZ1) project. We firstly present the specific features of GZ1 (Subsection IV-A), concerning the running of the project and available data. Then, we introduce two experts catalogues that allow for the evaluation of both amateur classifications and the proposed approach (Subsection IV-B). After this, we explain the experiments carried out to test our method (Subsection IV-C). Finally, we summarise and discuss the obtained results (Subsection IV-D).

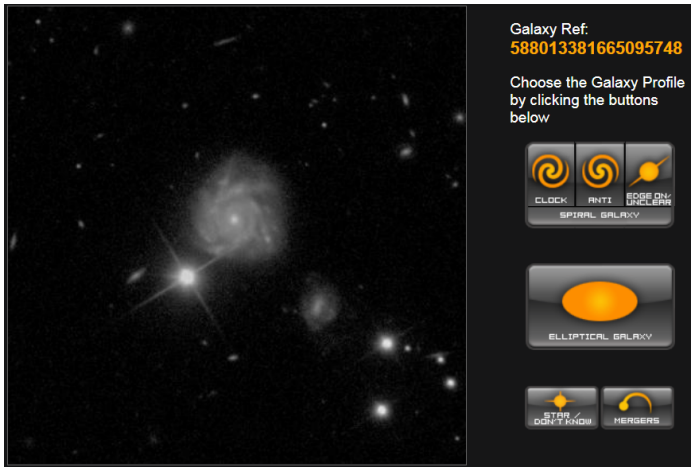

Fig. 2: View of GZ1 classification scheme, with the six categories used in the project: *Spiral* (three sub-categories at the top), *Elliptical* (middle), *Star / Don't Know* (bottom left), and *Merger* (bottom right).

A. Galaxy Zoo

The GZ1 project, launched in July 2007, brought the morphological classification of galaxies to a great number of amateurs engaged in making a significant contribution to the astrophysical research [16]. An application was made available on-line[3], and registered participants were able to classify galaxy images taken from the Sloan Digital Sky Survey[4], one of the main collections of astronomical images compiled to date. The project concentrated all efforts in helping astronomers disentangle the bimodality observed in galaxy morphologies, which divides the population between elliptical and spiral galaxies. The launch of the project was very successful, and after a period of six months over 100,000 volunteers had completed more than 40 million classifications for a sample of nearly 900,000 galaxy images [12].

In this project people were asked to classify galaxies into one of six categories: *Elliptical*, *Clockwise Spiral*, *Anti-clockwise Spiral*, *Edge-on Spiral*, *Star / Don't Know* and *Merger*, paying special attention to *Elliptical* and *Spiral* as the two main classes (Figure 2). Colour images of 423x423 pixels were shown to amateur participants, setting a universal image scaling for all objects in order to ensure that all classifications were made on a similar basis. Each galaxy ended up with a mean number of $\sim$38 independent classifications (votes), with a standard deviation of $\sim$14 votes. This data was conveniently analysed by a team of experts to evaluate the influence of different biases in the classification. This analysis resulted in a study completed by Bamford et al. [19] which developed a correction of the scores in favour of elliptical classifications. This debiasing of the scores was intended to prevent blurred

[3]The original GZ1 portal is maintained at http://zoo1.galaxyzoo.org.
[4]http://www.sdss.org

images of spiral galaxies to be classified as elliptical, for which the three spiral sub-types were added in a combined spiral score, which we will refer to as *Spiral* score from now on.

After the project had finished collecting votes in 2009, the GZ1 data was compiled in a set of csv files that was made publicly available[5]. These csv files contain the ID of the galaxy in the SDSS database, its location in the sky, total number of votes for the galaxy, original scores for all categories, and debiased scores for the two main categories: *Elliptical* and *Spiral*. Furthermore, there are classifications computed by the team of experts, using the debiased scores. These classifications, the so-called *GZ1 flags*, were computed applying a threshold of 0.8 over the debiased scores. This means that any galaxy with *Elliptical* or *Spiral* debiased score equal or greater than 0.8 were labelled as being elliptical or spiral, respectively, and *uncertain* in any other case. However, the debiasing needed an additional parameter[6] for its implementation, which was not available for the whole GZ1 dataset. As a result, the debiasing and GZ1 flags were only computed for a subset consisting of 667,944 galaxies, which we will refer to as GZ1 subset from now on.

B. Expert validation

Two expert catalogues were originally used by the GZ1 team in order to evaluate amateurs' performance [12]. On the one hand, the MOSES expert catalogue [20] contains 16,516 galaxies included in the GZ1 subset, and classified by professional astronomers as elliptical. On the other, the Longo expert catalogue [21] agglutinates 25,190 galaxies labelled as spiral by another set of experts and included in the GZ1 subset as well. Nonetheless, it is remarkable that there is an overlapping between both expert catalogues. This concerns 141 galaxies, which are removed for the consistency of results. After this correction, we consider the joint expert catalogue, which is composed of 41,424 galaxies from the GZ1 subset. We refer to this sample as validation subset. This subset plays a fundamental role as the available expert knowledge about the GZ1 project. It allows us to perform an expert validation as a ground truth needed for the testing of the experiments.

In what follows, we use two metrics for the validation of results: Accuracy (Acc) and Rejection Rate (RR). Acc tells about the proportion of proper classifications with respect to the total number of classified objects. The RR, instead, is computed taking the proportion of non-classified objects (*uncertain*) with respect to the sample size. Considering both measures, the validation subset permits an evaluation of GZ1 flags (Table I), which provides a benchmark for subsequent trials (Table II).

This form of expert validation implies the consideration of GZ1 classifications as a binary classification problem, with *Elliptical* and *Spiral* regarded as negative and positive classes, respectively. Under this view, *Merger* and *Star / Don't Know* categories are considered as secondary classes that, in fact, do not count with any form of expert validation. However, our approach aims to improve final classifications by leveraging information contained in DK votes.

[5]http://data.galaxyzoo.org

[6]This is the redshift, a fundamental measure in astrophysics that works as an indicator of the distance to the galaxy.

	MOSES	**Longo**	**Joint**
Present in GZ1 subset	16,375	25,049	41,424
Correctly flagged	4,181	20,385	24,566
Incorrectly flagged	1,040	26	1,066
Flagged as *uncertain*	11,154	4,638	15,792

TABLE I: Expert validation of GZ flags using MOSES (second column) and Longo (third column) expert catalogues separately and the joint catalogue (fourth column), after removing the 141 overlapped galaxies.

Accuracy	0.9584
Rejection Rate	0.3812

TABLE II: Evaluation of GZ1 flags, using the joint expert catalogue over the validation subset.

C. Experimental setting

We carry out two sets of experiments to test the adequacy of our proposed method. In the first set, we check the use of the IU to improve final classifications obtained from amateur votes, previously named as vote vector $\mathbf{N}$. To this aim, we firstly calculate the normalised score vector $\mathbf{Z}$ for each galaxy in the validation subset as shown in Equation 1. In GZ1, *Elliptical* and *Spiral*[7] categories constitute the main classes, covering the greatest part of the votes; *Merger* category works as the rare class, and *Star / Don't Know* category (DK votes) accounts for the MU.

After a normalisation of the debiased scores, we apply a series of thresholds in order to obtain final classifications: galaxies with *Elliptical* or *Spiral* normalised score greater or equal than the threshold are labelled as being elliptical or spiral, respectively. In any other case, the galaxy is annotated as *uncertain* and counts as not classified. We apply a set of thresholds over the scores, from 0.5 to 1.0 using 0.1 steps ([0.5-1.0]). With these cuts, we can check the trade-off between Acc and RR measures as we make more stringent the confidence over the data, that is, as we diminish the IU across the examples.

Figure 3 summarises the results of the first trial. We use the joint expert catalogue (Table I) to validate the results and calculate the Acc-RR measures. Each threshold in the [0.5-1.0] interval produces one Acc-RR point in the chart. We also evaluate the debiased scores before the normalisation in order to visualise the improvement provided by the normalisation of the scores. The set of points is presented in Table III as well.

The second set of experiments implements the use of MU. Before presenting the new trial with the shifted scores, we firstly explore the spread of MU within the validation subset. The mean number of DK votes is $\overline{n}_{DK} = 1.13$, with a standard deviation of $\sigma_{DK} = 1.83$. The greatest DK value within the

[7]Recall that we are referring to *Spiral* as the aggregation of original *Clockwise*, *Anti-clockwise* and *Edge-on* categories.

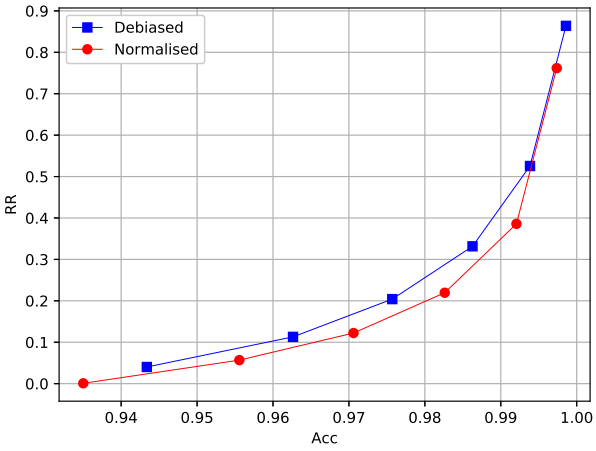

Fig. 3: Test using IU: the chart shows Acc-RR points generated by the application of [0.5-1.0] thresholds, from left to right, over debiased (blue squares) and normalised (red dots) scores. We use the joint expert catalogue as ground truth to validate the classifications.

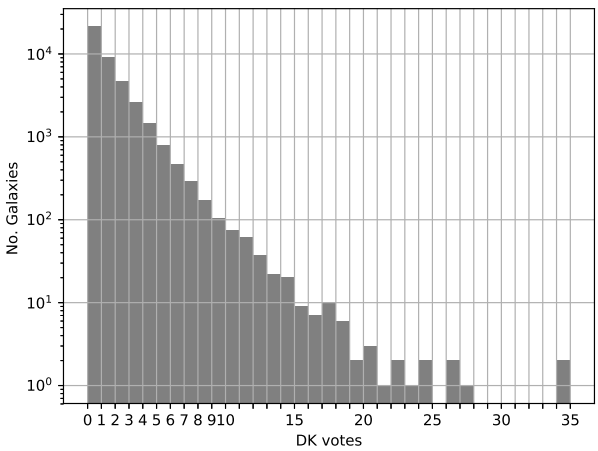

Fig. 4: Distribution of DK votes across the validation subset in logarithmic scale. The maximum value is 35, with $\overline{n}_{DK} = 1.13$ and $\sigma_{DK} = 1.83$.

	Debiased		Normalised	
Threshold	**Acc**	**RR**	**Acc**	**RR**
1.0	0.9986	0.8638	0.9974	0.7618
0.9	0.9938	0.5255	0.9921	0.3858
0.8	0.9863	0.3316	0.9826	0.2195
0.7	0.9757	0.2041	0.9706	0.1222
0.6	0.9626	0.1130	0.9556	0.0568
0.5	0.9434	0.0402	0.9350	0.0008

TABLE III: Set of Acc-RR points after the application of [0.5-1.0] thresholds, over debiased (left columns) and normalised (right columns) scores. We use the joint expert catalogue as ground truth to validate the classifications.

validation subset is 35, and there are 21,374 galaxies for which $n_{DK} = 0$ and consequently MU is zero as well. However, the spread of MU across the rest of the sample shows a continuous distribution throughout the validation subset (Figure 4).

Now we take the normalised scores previously obtained in the first trial and apply the shift presented in Equations 2 and 3, obtaining by this way the shifted score vector $\mathbf{W}$ for each galaxy in the validation subset. In this case study, we adopt the values $\alpha = -0.13$ and $\beta = 0.8$ (Equation 2), which are empirically found after running several tests with the original scores. Once again, we apply the same series of thresholds to get final classifications. As in the first trial, if the shifted scores do not reach the threshold, the galaxy is labelled as *uncertain* and counts as not classified.

Figure 5 presents the results of the second trial. As before, the validation is carried out using the joint expert catalogue so that each threshold in [0.5-1.0] generates one Acc-RR point in the chart. This time, we include the debiased and normalised scores along with the shifted scores. This set of points is also presented in Table IV.

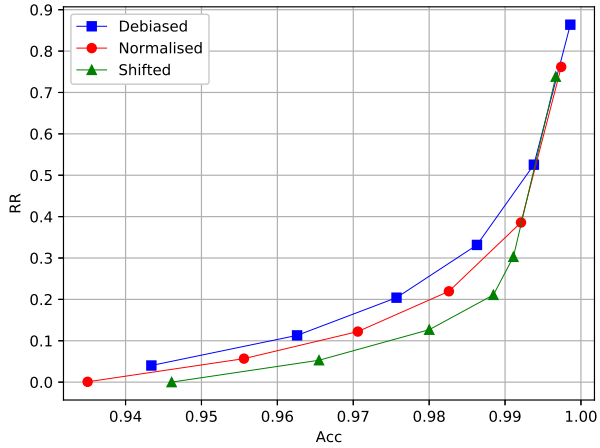

Fig. 5: Test using MU: the chart shows Acc-RR points generated by the application of [0.5-1.0] thresholds, from left to right, over debiased (blue squares), normalised (red dots), and shifted (green triangles) scores. We use the joint expert catalogue as ground truth to validate the classifications.

D. Discussion of results

Through the two sets of experiments, the new Acc-RR points provided represent a better trade-off with respect to those ones obtained from original scores. In the following, we extract the most representative results from the two sets of experiments, in accordance with the plots and tables showed above:

- The IU test (Figure 3) shows an overall improvement of the RR marks, considering Acc-RR points of equal threshold. This tendency is easier to observe in the tabulated values (Table III): RR measures obtained using normalised scores systematically outperform those ones

	Shifted	
Threshold	**Acc**	**RR**
1.0	0.9967	0.7382
0.9	0.9911	0.3028
0.8	0.9885	0.2109
0.7	0.9800	0.1267
0.6	0.9655	0.0529
0.5	0.9461	0.0

TABLE IV: Set of Acc-RR points after the application of [0.5-1.0] thresholds over shifted scores. We use the joint expert catalogue as ground truth to validate the classifications.

using debiased scores. This difference reaches the maximum with a 0.9 threshold, improving the RR in nearly a 14%, this is, classifying around 5,800 galaxies previously tagged as *uncertain*. The other great achievement is the RR obtained with the 0.5 threshold: this cut is equivalent to picking the greatest score. However, for 35 galaxies in the validation subset $\mathbf{X} = \mathbf{Z} = (0.5, 0.5)$, and IU takes the maximum. This first stage is unable to disentangle this disparity, which will be corrected using the MU in the second stage. The Acc values get worse with the normalisation of the scores, if we consider Acc-RR points of equal threshold. These differences are bigger as the threshold gets smaller. One possible explanation can be that the relaxation of the threshold introduces more noisy examples for which the IU outweighs amateur votes, that is, the normalisation lowers the scores so that both are too close to 0.5, and this turns random the final classifications. This issue is partially improved with the use of MU as well.

- The distribution of MU within the validation subset (Figure 4) calls for an employment of this information obtained from amateurs' clicks and never taken into consideration before. The wide variability throughout the data along with the fact that more than 48% of the galaxies hold DK votes, clearly indicate that the consideration of MU is able to boost final classifications, as it is effectively shown through the second set of experiments.
- The MU test (Figure 5), in contrast, outperforms both RR and Acc marks obtained with debiased and normalised scores for 0.5, 0.6, 0.7 and 0.8 thresholds. These marks represent the best trade-off in Acc-RR measures provided by the proposed approach. Nonetheless, the Acc-RR points for 0.9 and 1.0 thresholds do not obtain better marks with respect to the debiased and normalised scores. These results may indicate a worse behaviour of the approach as the IU diminishes, that is, with examples that hold a great consensus.

V. Conclusions and further work

In this paper, we have proposed a novel fuzzy-based use of widespread uncertainty within Citizen Science data. Its main achievement is to handle two different types of uncertainty prevalent to this sort of data: the first one is the so-called inherent uncertainty, and it is due to the lack of consensus across Citizen Science participants; the second one is referred to as measured uncertainty, and it is included within the set of amateur votes itself. Using these two measures, our method provides new score vectors that have shown a considerably improvement in final classifications. To test our approach, we have taken a representative project as case study: the Galaxy Zoo. Two sets of experiments addressing the two types of uncertainty, respectively, have demonstrated that the proposed approach is able to enhance the accuracy and rejection rate measures in classifications obtained by the application of a set of thresholds. This means better classifications in accordance with experts and a lower amount of *uncertain* galaxies.

As future work, we aim to extend this approach to other scenarios involving classification problems with more than two main classes. In such problems, the normalisation of the score vectors must generate another vector accounting for the inherent uncertainty, and not a single measure as occurs in binary classification. Therefore, the shifting of these vectors will entail more complex analyses. We also plan to study the use of the new score vectors presented here as input to a machine learning classifier. Finally, we aim to study the potential of traditional fuzzy approaches for the addition of information regarding either experts' or amateurs' individual performances.

References

[1] J. Cohn, Citizen science: Can volunteers do real research?, BioScience 58 (3) (2008) 192–197.

[2] N. Ball, R. Brunner, Data mining and machine learning in astronomy, International Journal of Modern Physics D 19 (7) (2010) 1049–1106.

[3] F. Candido dos Reis, S. Lynn, H. Ali, D. Eccles, A. Hanby, E. Provenzano, C. et al., Crowdsourcing the general public for large scale molecular pathology studies in cancer, EBioMedicine 2 (7) (2015) 681–689.

[4] M. Banerji, O. Lahav, C. Lintott, F. Abdalla, K. Schawinski, S. Bamford, D. Andreescu, P. Murray, M. Raddick, A. Slosar, A. Szalay, D. Thomas, J. Vandenberg, Galaxy zoo: Reproducing galaxy morphologies via machine learning, Monthly Notices of the Royal Astronomical Society 406 (1) (2010) 342–353.

[5] S. Dieleman, K. Willett, J. Dambre, Rotation-invariant convolutional neural networks for galaxy morphology prediction, Monthly Notices of the Royal Astronomical Society 450 (2) (2015) 1441–1459.

[6] M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, K. Crowston, A. Katsaggelos, S. Larson, T. Lee, C. Lintott, T. Littenberg, A. Lundgren, C. Osterlund, J. Smith, L. Trouille, V. Kalogera, Gravity spy: Integrating advanced ligo detector characterization, machine learning, and citizen science, Classical and Quantum Gravity 34 (6) (2017) 64003–64025.

[7] H. Show, Rise of the citizen scientist, Nature 524 (2015) 265.

[8] R. Bonney, J. Shirk, T. Phillips, A. Wiggins, H. Ballard, A. Miller-Rushing, J. Parrish, Next steps for citizen science, Science 343 (6178) (2014) 1436–1437.

[9] M. Kosmala, A. Wiggins, A. Swanson, B. Simmons, Assessing data quality in citizen science, Frontiers in Ecology and the Environment 14 (10) (2016) 551–560.

[10] J. Garibaldi, T. Ozen, Uncertain fuzzy reasoning: A case study in modelling expert decision making, IEEE Transactions on Fuzzy Systems 15 (1) (2007) 16–30.

[11] E. Madi, J. Garibaldi, C. Wagner, Exploring the use of type-2 fuzzy sets in multi-criteria decision making based on topsis, FUZZ-IEEE 2017 – 2017 IEEE International Conference on Fuzzy Systems (2017) 1–6.

[12] C. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M. Raddick, R. Nichol, A. Szalay, D. Andreescu, P. Murray, J. Vandenberg, Galaxy zoo: Morphologies derived from visual inspection of galaxies from the sloan digital sky survey, Monthly Notices of the Royal Astronomical Society 389 (3) (2008) 1179–1189.

[13] R. Simpson, K. Page, D. De Roure, Zooniverse: Observing the world's largest citizen science platform, WWW 2014 – 23rd International Conference on World Wide Web (2014) 1049–1054.

[14] J. Silvertown, A new dawn for citizen science, Trends in Ecology and Evolution 24 (9) (2009) 467–471.

[15] D. Bonter, C. Cooper, Data validation in citizen science: A case study from project feederwatch, Frontiers in Ecology and the Environment 10 (6) (2012) 305–307.

[16] L. Fortson, K. Masters, R. Nichol, K. Borne, E. Edmondson, C. Lintott, J. Raddick, K. Schawinski, J. Wallin, Galaxy zoo: Morphological classification and citizen science, Machine Learning and Data Mining for Astronomy 11 (2012) 118–125.

[17] K. Crowston, C. Østerlund, T. K. Lee, Blending machine and human learning processes, HICSS 2017 – 50th Hawaii International Conference on System Sciences.

[18] F. Herrera, E. Herrera-Viedma, Linguistic decision analysis: Steps for solving decision problems under linguistic information, Fuzzy Sets and Systems 115 (1) (2000) 67–82.

[19] S. Bamford, R. Nichol, I. Baldry, K. Land, C. Lintott, K. Schawinski, A. Slosar, A. Szalay, D. Thomas, M. Torki, D. Andreescu, E. Edmondson, C. Miller, P. Murray, M. Raddick, J. Vandenberg, Galaxy zoo: The dependence of morphology and colour on environment, Monthly Notices of the Royal Astronomical Society 393 (4) (2009) 1324–1352.

[20] K. Schawinski, D. Thomas, M. Sarzi, C. Maraston, S. Kaviraj, S.-J. Joo, S. Yi, J. Silk, Observational evidence for agn feedback in early-type galaxies, Monthly Notices of the Royal Astronomical Society 382 (4) (2007) 1415–1431.

[21] M. Longo, Detection of a dipole in the handedness of spiral galaxies with redshifts z 0.04, Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics 699 (4) (2011) 224–229.