WIP: Citizen Science Tools with Machine Learning as a Pathway to Engage High School Students in Research

Fahim Hasan Khan
Computer Science and Engineering
University of California, Santa Cruz
fkhan4@ucsc.edu

Emily Lovell
Computer Science and Engineering
University of California, Santa Cruz
emme@soe.ucsc.edu

Akila de Silva
Computer Science
San Francisco State University
desilva@sfsu.edu

Gregory Dusek
NOAA National Ocean Service
gregory.dusek@noaa.gov

James Davis
Computer Science and Engineering
University of California, Santa Cruz
davis@cs.ucsc.edu

Alex Pang
Computer Science and Engineering
University of California, Santa Cruz
pang@soe.ucsc.edu

***Abstract*—This research-to-practice WIP paper describes an approach to engage high school students in research through the utilization of citizen science tools embedded with Machine Learning (ML) models. In the context of fostering early engagement in scientific research among high school students, this paper explores the integration of citizen science and ML using SmartCS, an existing platform for creating citizen science smartphone applications. The process requires no prior programming knowledge, making it accessible to a broad range of students. For our approach, a group of high school students participated in a two-month-long summer research program, where they were introduced to the principles of citizen science as a method for data collection across diverse scientific projects from different research domains. The program's initial task involved students in the conceptualization of a citizen science project, adopted based on a thorough literature review, followed by the practical task of developing a smartphone application for data collection and educational purposes. Students either created new datasets or curated existing ones to train lightweight ML models for computer vision tasks, specifically focused on providing visual guidance within these mobile apps. The final task involved deploying these applications for public use and collecting user feedback. Our experience suggests that this approach not only enabled students to learn aspects of computer science and engineering, particularly in the area of ML model training and mobile application software development, but also allowed them to experience firsthand the significant role citizen science can play in collecting and analyzing scientific data.**

I. Introduction

In recent years, there has been a growing emphasis on incorporating research methodologies and scientific thinking into early education to enrich high school students' learning experiences and better prepare them for future academic and professional challenges [1]. This engagement fosters a passion for science and technology while enhancing critical thinking, problem-solving skills, and a sense of contribution [2]. This WIP (work-in-progress) paper introduces an approach that combines citizen science applications with machine learning (ML) to involve high school students in meaningful research.

Citizen science enables public participation in research projects, typically involving data collection through smartphone apps, which professional researchers then analyze [3]. For instance, apps can be developed to collect images of coastal sea life for ocean science or leaf images for botanists. These apps can be enhanced with ML to provide automated guidance for complex data collection tasks [4], [5]. By leveraging a platform for the codeless creation of ML-enhanced apps, high school students can participate in impactful research projects without needing advanced technical skills.

This work focuses on developing citizen science apps for collecting and analyzing visual data (images and videos), utilizing computer vision ML models [6]. This hands-on approach offers students socially relevant and impactful exposure to ML. While recent advancements in large language models, such as Generative Pre-trained Transformers (GPTs), have been applied in citizen science [7], applying ML for computer vision-based guidance is more challenging due to the subjectivity of visual data. Object detection or identification, a key technique in computer vision [8], is especially useful in citizen science applications that provide visual cues to participants. Using the codeless platform SmartCS [9], we created citizen science apps with object detection capabilities, allowing high school students to explore ML through the application of computer vision in research.

This paper investigates the following research questions (RQs):

1) RQ1: How does integrating ML within citizen science tools influence high school students' engagement and interest in scientific research?
2) RQ2: How effective is using a codeless platform in teaching high school students complex concepts such as data curation, ML, and mobile application development?

3) RQ3: What learning outcomes are observed among high school students who develop ML-powered mobile applications through the SmartCS platform?
4) RQ4: What impact do student-developed ML-powered apps have on public participation in citizen science projects?

II. Related Work

A. Citizen Science in High School Education

In the past, there have been various efforts to integrate citizen science into the learning process at the high school level [10]. However, it is much less prevalent than in higher education [11]. For high school students, citizen science offers learning opportunities and engagement across various scientific fields [12]. It also serves as an early introduction to and motivation for STEM careers [13]. Its benefits have been demonstrated for students as young as those in the fifth [14] and eighth grades [15]. Also, there have been initiatives for driving innovation through project-based learning for social good, participated in by middle and high school students [16]. This project aims to further bridge the gap between high school education and advanced scientific research by leveraging the power of citizen science.

B. ML in High School Education

Teaching fundamental AI (Artificial Intelligence) concepts and techniques, including ML as a sub-field, has traditionally been confined to higher education [17]. Recently, computing education has begun to be included in high school curricula worldwide, incorporating some advanced topics [18]. For instance, [19] introduced a sandbox methodology for teaching computing-based data science to high school students. However, computer science course content at this educational level rarely covers AI or ML [20]. In the last year, though, there has been an influx of GPT-4-based applications for education, including those aimed at high school students [21]. Even though opportunities for high school students to explore ML and its societal implications have been investigated, approaches for teaching them about ML and its application in research remain very limited [22]; this is something we strive to address with our work.

C. ML and Citizen Science

Limited but successful applications of ML have been observed in the field of Citizen Science in the past, including mobile apps like iNaturalist [5], PlantNet [23], and LeafSnap [24]. However, the untapped potential of using ML in citizen science remains vastly unexplored [9]. Echeverria et al. [25] demonstrated that the citizen science platform iNaturalist is a valuable tool for carrying out collaborative projects in secondary education. Our work aims to build on past successes by integrating ML to improve the accessibility of citizen science projects. This could potentially lead to a new paradigm in high school educational methodologies, making learning more interactive and engaging.

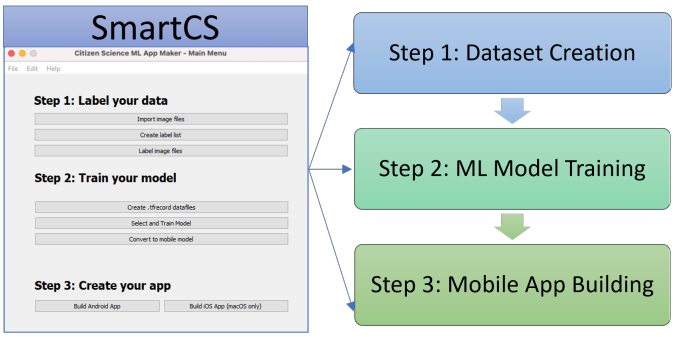

Fig. 1. The graphical user interface of SmartCS facilitates the three steps required to create an ML-powered citizen science mobile app.

III. Methodology and Research Setting

For this research, we used both qualitative and quantitative methods to investigate the RQs, with a strong focus on qualitative analysis. The data was collected through participant observation, informal interviews, feedback analysis, and using some basic quantitative measures like project completion rate, usability tests, and Likert scales.

A. Participant Selection

The high school students were recruited through a competitive selection process by an annual summer science internship program offered by an R1 institution. A research mentor, typically a graduate student, provides the students with guidance throughout the project. It was clearly communicated in the participant call that no programming skills were required for participation in the project. We enrolled 12 students (seven male, five female) aged 14 to 17, from diverse demographics, including the USA and India. In addition, data was collected on each participant's prior experience with technology, educational background, and research interest.

B. Structure

Our initiative provided a comprehensive research experience in three phases: (1) conceptualization of citizen science projects, (2) development, and (3) deployment of mobile apps. Table I outlines the learning goals and tools for each phase.

1) Conceptualization Phase: In this phase, students explore the application of ML to citizen science and identify potential research areas through an open-ended approach. Guided by mentors, students define the scope and objectives of their projects—such as designing an app to detect whether a given object is recyclable. The mentor ensured that the projects were feasible to develop within the given two-month timeframe and that students were reviewing relevant literature while ensuring they received sufficient information about the tasks at hand. They learned how to conceptualize a science project by reviewing current literature, identifying a research problem, and proposing a solution that integrates computer vision, ML, and citizen science. Thus, this phase introduced them to research methodology.

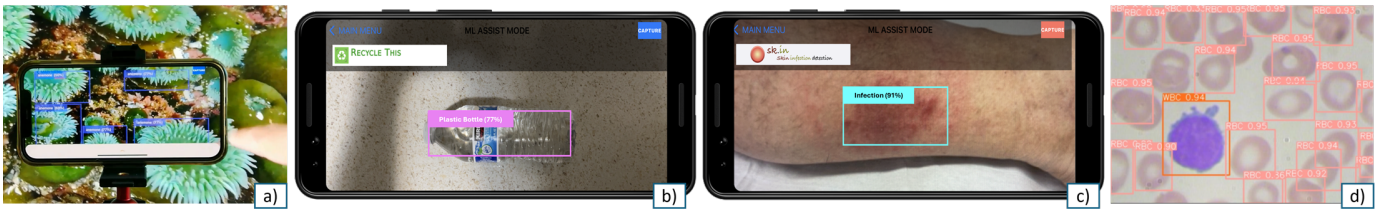

Fig. 2. Some of the apps created by the high school students include (a) Tidepool species identification, (b) Recyclable object detection, (c) Skin infection identification, and (d) Blood cell type identification.

2) Development Phase: The primary tool for developing ML-powered applications was SmartCS App Studio and its supporting tools [9]. This phase, guided by SmartCS, involves three steps: dataset creation, ML model training, and mobile app building (Figure 1). This phase includes training on ML basics, data collection and processing, and app development. Examples of student-created apps are shown in Figure 2.

a) Dataset Creation: Students gathered and labeled data, ensuring it was appropriately formatted for ML model training. For instance, the tidepool species identification app involved collecting many clear images of tidepool species suitable for computer-vision-assisted identification. This step helped them understand the importance of quality data and how it directly impacts the effectiveness of ML models. Data was sourced from existing (often multiple) public datasets or collected manually, depending on availability. Open-source tools, such as LabelImg and AnyLabeling, were used to label images. If any conversion of dataset formats was needed, the free tool Roboflow was used.

b) ML Model Training: This step teaches students the fundamentals of ML algorithms and the model training process. Participants used prepared datasets to train at least one lightweight ML model suitable for mobile devices, selecting from SSD-Mobilenet, EfficientDet, and YOLO models [8]. The training was conducted on a local or remote computer using tools like Google Colab, Amazon Web Services, and Ultralytics Hub. This experience introduced students to ML concepts in computer vision.

c) Mobile App Building: Students integrated their trained ML models into an iOS and/or Android mobile app using pre-developed templates provided by SmartCS. The learning objective here was to comprehend how ML models can be applied in real-world applications through mobile app development. This step also served as students' introduction to software engineering. They either used Xcode for iOS app creation or Android Studio for Android app creation (Figure 2). In this step, students' visions for their apps came to life. For example, an app could allow a user to identify a tidepool species in real-time using their smartphone camera, with the image then uploaded for further research use.

3) Deployment Phase: The mobile apps were distributed among beta testers, and students collected and analyzed user feedback. This served as their introduction to software testing, teaching them how to gather user feedback and use it to refine their projects. They also gained exposure to the human-computer interaction process.

IV. Preliminary Results and Discussion

Here, we reflect on our preliminary observations as guided by the RQs. We evaluated outcomes through observations of student learning during the program, informal student comments, measures of project completion success, and beta testing of two student apps.

1) RQ1: How does integrating ML within citizen science tools influence high school students' engagement and interest in scientific research?: We observed that integrating ML within citizen science tools significantly boosts high school students' engagement and interest in scientific research, perhaps by making learning more interactive and impactful. We hypothesize that the ML-enhanced app's real-time feedback using object identification to enable learning and guide data collection makes the scientific process more exciting and understandable. This unique approach to designing science projects may help maintain the curiosity and excitement of the students creating the apps. Moreover, hands-on application of cutting-edge technology like ML exposes them to modern scientific methods, potentially sparking a lasting interest in STEM careers. By contributing to real-world research through state-of-the-art yet user-friendly ML tools (as listed in Table I), students can see the practical impact of their work, enhancing their sense of contribution and the relevance of their educational activities.

2) RQ2: How effective is using a codeless platform in teaching high school students complex concepts such as data curation, ML, and mobile application development?: We hoped that using a codeless platform to teach high school students complex concepts would enable them to rapidly plan and engage in meaningful scientific research without requiring extensive background knowledge. The number of functional applications students have created, as summarized in Table II, suggests this may indeed be a successful approach. This direct engagement with technology may enhance students' understanding through experiential learning. By making these tools accessible, students from diverse backgrounds can develop crucial digital age skills such as data curation, understanding ML algorithms, and gaining app development experience.

Qualitative feedback from participants also underscores the program's value. For instance, Student S7 remarked, "*I really enjoyed the program as it was beginner-friendly. Having never done any ML work previously, it showed the benefits and*

TABLE I
SUMMARY OF LEARNING ACTIVITIES BY PHASES AND STEPS

Phase	Step	Learning Goals	Tools Learned
Conceptualization	-	Study literature and state-of-the-art to conceptualize the idea of a project	Google Scholar
Development	Dataset Creation	Creating a dataset or repurposing a public dataset	SmartCS, LabelImg, AnyLabeling, Roboflow
	ML Model Training	Training an ML model using local PC/cloud computing	SmartCS, Google Colab, AWS, Ultralytics Hub
	Mobile App Building	Building an app for iOS and/or Android phones	SmartCS, Xcode, Android Studio
Deployment	-	Recruit users, distribute the apps, and collect user feedback	TestFlight, Google Drive, Google Forms

TABLE II
LIST OF PROJECTS CREATED BY THE HIGH SCHOOL STUDENTS

Student	Project	ML Model	Phone App	Publi-cation
S1	Tidepool species identification	Yes	Yes	-
S2	Recyclable object detection	Yes	Yes	Yes
S3	Skin infection identification	Yes	Yes	-
S4	Road vehicle type recognition	Yes	Yes	-
S5	Plant disease detection	Yes	-	-
S6	Beach debris identification	Yes	Yes	-
S7	Building architecture identification	Yes	Yes	-
S8	Fingernail conditions assessment	Yes	Yes	-
S9	Blood cell type identification	Yes	Yes	Yes
S10	Ultrasound images type classification	Yes	-	-
S11, S12	Differentiate between seals and sea lions	Yes	Yes	-

possibilities while making it easy to label, create, and turn models into apps." S9 commented, "*The app creation platform expedited the process of creating the ML app and allowed me to easily deploy the ML models trained to perform real-time detection. Furthermore, the app creation platform allowed us to control various parameters, which would improve the model performance.*" We find it especially noteworthy that students effectively used ML terminology to describe their experiences.

3) RQ3: What learning outcomes are observed among high school students who develop ML-powered mobile applications through the SmartCS platform?: High school students who participated in developing ML-powered mobile applications using the SmartCS platform were expected to achieve several key learning goals and learn some crucial software tools, as listed in Table I.

For example, using AnyLabeling, which allows data labeling with AI support from YOLO and Segment Anything, students can learn how ML models support real-life computer vision tasks. Additionally, the process fosters critical thinking, problem-solving, collaboration, and communication skills. These competencies enhance academic aptitude and prepare students for future STEM careers.

The apps successfully created by the students (Table II) suggest that students have likely met the learning goals, including having developed the skills to use the tools outlined in Table I. Notably, two students published their work at IEEE conferences following the two-month program [26], [27]. Publishing was not an intended goal of our initiative, as it is uncommon for high school students to publish, so this is especially suggestive of a transformative learning experience. This highlights both the long-term positive impact of our initiative and the potential of high school students to contribute to scientific research. Additionally, it is notable that five out of twelve students who have since graduated high school have already enrolled in STEM undergraduate programs.

4) RQ4: What impact do student-developed ML-powered apps have on public participation in citizen science projects?: To explore this RQ, we pilot-tested the usability of two selected apps. This usability testing aimed to demonstrate the effectiveness of ML-based object detection apps in helping participants capture useful data. We recruited 20 participants, consisting of 10 males and 10 females, aged between 18 to 54. They were asked to use student-created apps to conduct (1) tidepool species identification and (2) recyclable object detection. We conducted this test in a lab, which enabled us to determine the participants' success rates and collect their feedback.

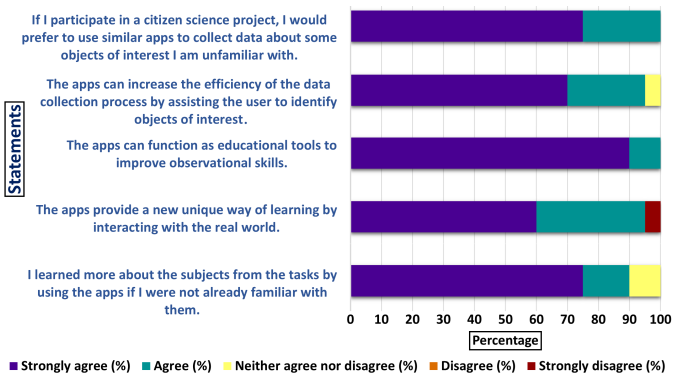

Fig. 3. Summary of feedback from users on their experience of using the apps. Here, in the statements, "the apps" refer to the two selected apps created by the students.

Feedback from app users (Figure 3) indicates that ML guidance improved participants' experiences. They acknowledged the positive impact of ML-powered apps on citizen science, highlighting the object identification feature as helpful for data collection. Additionally, participants viewed these apps as educational tools that provide interactive learning experiences, suggesting that ML-powered apps can significantly enhance public participation in citizen science by making data collection more accessible and engaging.

V. Conclusion and Future Work

This paper demonstrates the early effectiveness of integrating citizen science and ML to engage high school students in scientific research. A codeless platform like SmartCS has proven to be a valuable tool, enabling students to create meaningful projects that contribute to both their educational growth and the broader scientific community. The broader impact of this work includes fostering early interest in STEM and enhancing public engagement in scientific research.

In future work, we will formally evaluate each research question through rigorous user studies. We also plan to expand the scope of citizen science projects to include various domains and ML applications beyond object identification to achieve scalability. Additionally, we intend to investigate the effectiveness of this approach with younger students, including those in middle and elementary schools.

Acknowledgements

This work is partially funded by grants from SECOORA (NOAA award NA20NOS0120220), the USCRP (Sea Grant NA23OAR4170121), and the UCSC Center for Coastal Climate Resilience. The contents, findings, and conclusions are those of the authors and do not necessarily reflect the views, positions, or policies of the funding agencies or the government. No official endorsement should be inferred. We also acknowledge the Cloud Credits for Research support from Google Cloud and AWS.

This research was approved by the Office of Research Compliance Administration at UCSC, with informed consent obtained from all participants.

References

[1] Maarten Van Mechelen, Rachel Charlotte Smith, Marie-Monique Schaper, Mariana Tamashiro, Karl-Emil Bilstrup, Mille Lunding, Marianne Graves Petersen, and Ole Sejer Iversen. Emerging technologies in k–12 education: A future hci research agenda. *ACM Transactions on Computer-Human Interaction*, 30(3):1–40, 2023.

[2] Jennifer Tsan, David Weintrop, Donna Eatinger, and Diana Franklin. Learner ideas and interests expressed in open-ended projects in a middle school computer science curriculum. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 820–826, 2023.

[3] Christopher Kullenberg and Dick Kasperowski. What is citizen science?–a scientometric meta-analysis. *PloS one*, 11(1):e0147152, 2016.

[4] Fahim Hasan Khan, Akila de Silva, Gregory Dusek, James Davis, and Alex Pang. Authoring platform for mobile citizen science apps with client-side ml. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, pages 89–94, 2021.

[5] Jill Nugent. inaturalist: citizen science for 21st-century naturalists. *Science Scope*, 41(7):12–15, 2018.

[6] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Nature, 2022.

[7] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[8] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[9] Fahim Hasan Khan, Akila de Silva, Gregory Dusek, James Davis, and Alex Pang. Smartcs: Enabling the creation of machine learning–powered computer vision mobile apps for citizen science applications without coding. *Citizen Science: Theory and Practice*, 9(1), 2024.

[10] Harsh R Shah and Luis R Martinez. Current approaches in implementing citizen science in the classroom. *Journal of microbiology & biology education*, 17(1):17–22, 2016.

[11] Caterina Solé, Digna Couso, and María Isabel Hernández. Citizen science in schools: A systematic literature review. *International Journal of Science Education, Part B*, pages 1–17, 2023.

[12] Julia Kelemen-Finan, Martin Scheuch, and Silvia Winter. Contributions from citizen science to science education: an examination of a biodiversity citizen science project with schools in central europe. *International Journal of Science Education*, 40(17):2078–2098, 2018.

[13] Suzanne E Hiller and Anastasia Kitsantas. The effect of a horseshoe crab citizen science program on middle school student science performance and stem career motivation. *School Science and Mathematics*, 114(6):302–311, 2014.

[14] Ruth Kermish-Allen, Karen Peterman, and Christine Bevc. The utility of citizen science projects in k-5 schools: measures of community engagement and student impacts. *Cultural Studies of Science Education*, 14(3):627–641, 2019.

[15] Kathryn Paige, Robert Hattam, and Christopher B Daniels. Two models for implementing citizen science projects in middle school. *The Journal of Educational Enquiry*, 14(2), 2015.

[16] Gayathri Manikutty, Sreejith Sasidharan, and Bhavani Rao. Driving innovation through project based learning: A pre-university steam for social good initiative. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2022.

[17] Algeir P Sampaio, Paulo CMA Farias, and Roberto A Bittencourt. A case study of using machine learning in k-12 education. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2023.

[18] Gurmeher Kaur, Kris Jordan, and Jasleen Kaur. Using foundational cs1 curricula for middle school & early high school computer programming education. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 827–833, 2023.

[19] Justice T Walker, Amanda Barany, Alex Acquah, Sayed Mohsin Reza, Alan Barrera, Karen Del Rio Guzman, and Michael A Johnson. Coding like a data miner: A sandbox approach to computing-based data science for high school student learning. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–5. IEEE, 2023.

[20] Peter Hubwieser, Michal Armoni, and Michail N Giannakos. How to implement rigorous computer science education in k-12 schools? some answers and many questions. *ACM Transactions on Computing Education (TOCE)*, 15(2):1–12, 2015.

[21] Gloria Ashiya Katuka, Yvonika Auguste, Yukyeong Song, Xiaoyi Tian, Amit Kumar, Mehmet Celepkolu, Kristy Elizabeth Boyer, Joanne Barrett, Maya Israel, and Tom McKlin. A summer camp experience to engage middle school learners in ai through conversational app development. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, pages 813–819, 2023.

[22] Magnus Høholt Kaspersen, Karl-Emil Kjær Bilstrup, Maarten Van Mechelen, Arthur Hjort, Niels Olof Bouvin, and Marianne Graves Petersen. High school students exploring machine learning and its societal implications: Opportunities and challenges. *International Journal of Child-Computer Interaction*, 34:100539, 2022.

[23] Hervé Goëau, Pierre Bonnet, Alexis Joly, Vera Bakić, Julien Barbe, Itheri Yahiaoui, Souheil Selmi, Jennifer Carré, Daniel Barthélémy, Nozha Boujemaa, et al. Pl@ ntnet mobile app. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 423–424, 2013.

[24] Neeraj Kumar, Peter N Belhumeur, Arijit Biswas, David W Jacobs, W John Kress, Ida C Lopez, and João VB Soares. Leafsnap: A computer vision system for automatic plant species identification. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, pages 502–516. Springer, 2012.

[25] Andres Echeverria, Idoia Ariz, Judit Moreno, Javier Peralta, and Esther M Gonzalez. Learning plant biodiversity in nature: The use of the citizen–science platform inaturalist as a collaborative tool in secondary education. *Sustainability*, 13(2):735, 2021.

[26] Nihar Jain and Fahim Hasan Khan. Blood cell detection using deep learning on mobile platforms. In *2023 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1289–1293. IEEE, 2023.

[27] Chelsea Yeh and Fahim Hasan Khan. Citizen science mobile apps with machine learning for recyclable objects. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1539–1542. IEEE, 2022.