Design and Implementation of a Three-Layer Backpropagation Neural Network for Multi-Output Regression in Citizen-Science Impact Assessment

Luigi Ceccaroni *, Lyle Visa, Iain Visa

Posted Date: 25 March 2026

doi: 10.20944/preprints202603.2068.v1

Keywords: neural network; multiclass classification; backpropagation; citizen science; input features; impact assessment

 doi:10.20944/preprints202603.2068.v1

Article

Design and Implementation of a Three-Layer Backpropagation Neural Network for Multi-Output Regression in Citizen-Science Impact Assessment

Luigi Ceccaroni [1,*], Lyle Visa [2] and Iain Visa [3]

[1] Earthwatch, 102-104 St Aldate's, Oxford OX1 1BT, UK
[2] University of Manchester, Oxford Rd, Manchester M13 9PL, UK
[3] University of Barcelona, Gran Via de les Corts Catalanes, 585, Eixample, 08007 Barcelona, Spain
* Correspondence: lceccaroni@earthwatch.org.uk

Abstract

Measuring the impact of citizen-science projects is hard because inputs are heterogeneous, mostly categorical, and sparse. We present Alquimics, a compact supervised neural network trained on one-hot project descriptors to predict impact across five domains (Environment, Economy, Governance, Science, and Society). Each project is encoded as a binary vector of length 4,460 (223 questions × 20 options, flattened). The network employs a 4,460–42–5 topology with logistic activations throughout; labels consist of five continuous targets in [0, 1] obtained by scaling expert domain scores in [1, 42]. We implement L2-regularised training in Octave using fmincg with MaxIter = 10 and lambda = 0.07. We document the entire data pipeline, objective, and implementation, provide a minimal reproducible script, and discuss limitations arising from the small dataset (n = 9 projects). This establishes a transparent baseline that complements rule-based scoring and can be expanded as more labelled projects become available.

Keywords: neural network; multiclass classification; backpropagation; citizen science; input features; impact assessment

1. Introduction

1.1. Background and Motivation

Citizen science has expanded dramatically across environmental monitoring, biodiversity, astronomy, and beyond (Bonney et al., 2014). Yet despite widespread adoption, quantitative evidence of its impact often lags significantly behind its growth. Assessing this impact is inherently challenging: it involves complex interactions among numerous factors spanning participation patterns, data practices, project design, and outcomes, rarely captured by simple rule-based systems alone.

1.2. The Challenge of Impact Assessment

Traditional rule-based scoring systems offer interpretability, allowing researchers to understand why a given assessment was reached. However, they struggle with the nonlinear interactions and subtle patterns that characterise real-world citizen-science projects. Machine learning approaches can complement these deterministic methods by identifying signals buried in numerous weak indicators. This capability is essential when a comprehensive impact assessment requires consideration of hundreds of input features that capture the full complexity of project design and implementation.

 doi:10.20944/preprints202603.2068.v1

1.3. The MICS Framework and Alquimics

The EU-funded MICS project (Monitoring for Impact in Citizen Science) (Parkinson et al., 2022; Sprinks et al., 2021) was established to address this assessment challenge through a structured, comprehensive approach. MICS collects detailed project descriptors via a battery of closed-ended questions spanning project design, participation dynamics, data practices, and outcomes, such as the following one (see also Figure 1):

Did the project generate related new projects?

- Yes
- It's in progress (for example, a project proposal has been written)
- No
- I don't know

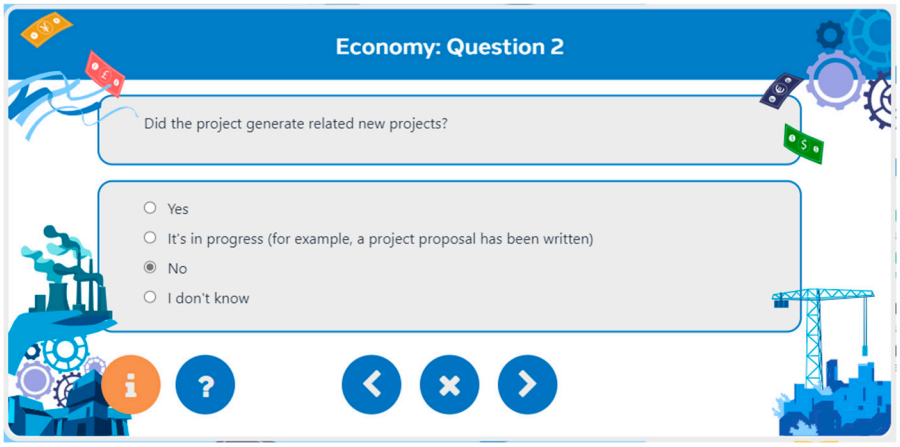

Figure 1. *Example of a question corresponding to an input feature in MICS.*

Five impact domains frame the assessment: Environment, Economy, Governance, Science, and Society. This paper describes Alquimics, a feedforward neural network developed within MICS to map project descriptors (captured via 223 structured input features) to the five impact domains. Alquimics is integrated into MICS's open-source web application, alongside a rule-based recommendation engine that generates personalised suggestions for improving impact.

1.4. Methodological Approach

The development of MICS and Alquimics has required navigating a fundamental design choice: which platform components should rely on handcrafted, explicit rules versus machine-learning models. Handcrafting (the traditional approach of writing extensive rule sets) provides transparency and domain control. Machine learning, conversely, enables systems to automatically discover patterns in data, making it particularly well-suited to classification tasks, where neural networks can uncover relationships in noisy, multidimensional data (Sarker, 2021). The MICS approach combines both: Alquimics employs supervised neural-network learning to capture complex patterns in project-impact relationships, while the broader MICS platform retains rule-based components to support interpretability and recommendations from *interested parties and actors* (IPAs) (Villena et al., 2011).

1.5. Research Scope and Contribution

This paper documents the dataset, model architecture, training procedure, and validation results for Alquimics. The primary innovation lies in applying neural networks to the most extensive feature set assembled to date for citizen-science impact assessment. The training dataset comprised nine complete instances (citizen-science projects) characterised by 223 input features. The resulting model represents an early, substantial step toward automated impact quantification in citizen science. We

 doi:10.20944/preprints202603.2068.v1

position this work within a pragmatic assessment pipeline that recognises both the strengths of machine learning for pattern discovery and the continued necessity of human-readable interpretability for IPA engagement.

2. Materials and Methods

2.1. Dataset and Labels

Source. Inputs are answers to **223** closed questions in the MICS platform's self-assessment. Each question is represented by a **20-bin** one-hot vector. 20 represents the maximum number of possible answers for a question.

Instances. We use **nine** projects with complete responses. For each project, we have five domain impact-scores on an ordinal scale from 1 to 42 (Environment, Economy, Governance, Science, and Society).

Label scaling. Targets are normalised to **[0, 1]** by dividing by 42 for training; predictions are rescaled to **[1, 42]** for reporting.

2.2. Feature Engineering

All inputs are already provided as one-hot matrices of shape **223 × 20** per project. We **flatten** each matrix column-major into a single **4,460-length** vector, concatenate a bias unit during training, and keep values binary. No imputation is required because only fully answered projects are included.

2.3. Model Architecture and Training

Topology. Fully connected **4,460–42–5** multilayer perceptron with logistic activations in the hidden and output layers (see Figure 2).

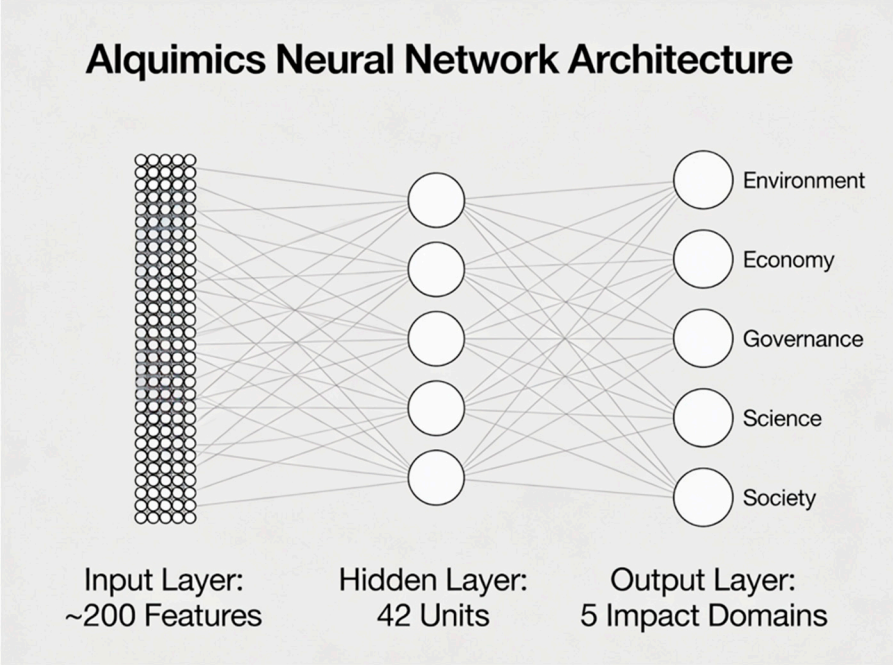

Figure 2. *Structure of the Alquimics neural network.*

Loss. Sum of five logistic cross-entropy terms against the five scaled targets, plus L2 weight penalty.

Optimiser. fmincg (conjugate-gradient) with **MaxIter = 10**.

Regularisation. lambda = 0.07 during training; higher values (e.g., **lambda = 1**) are used only for diagnostic cost checks.

Initialisation. Small random weights; bias units handled explicitly.

 doi:10.20944/preprints202603.2068.v1

2.4. Problem Formulation

This is a **multi-output regression with logistic heads**: each of the five outputs independently models the scaled score for one domain. This avoids forcing trade-offs across domains and accommodates use cases in which domain scores are not mutually exclusive (Xu et al., 2019).

2.5. Evaluation Protocol and Metrics

Because only nine projects are available, model performance is tracked during training by computing the loss function and the *root mean squared error* (RMSE) for each domain. The RMSE values are converted back to the original 1–42 scale of the expert scores for more straightforward interpretation. For deployment-ready estimates, the recommended protocol is *leave-one-project-out cross-validation* (LOOCV) with per-domain RMSE and *mean absolute error* (MAE). Confusion matrices are not applicable because the targets are continuous.

Reporting. A script writes the predicted 5-vector for each run to impact.csv after rescaling to **[1, 42]**.

2.6. Implementation Details and Code Structure

The Octave implementation follows a compact, didactic layout. Entry points are mics.m and trainingMics.m. Core modules implement forward and backward passes with explicit handling of bias terms. File roles are as follows:

- mics.m: loads data, launches training with defaults, writes predictions to impact.csv.
- trainingMics.m: initialises weights, sets regularisation and optimiser options, calls fmincg.
- nnCostFunction.m: forward pass, logistic activations, vectorised cost with L2 penalty, backprop gradients.
- predict.m: forward pass for inference only.
- sigmoid.m, sigmoidGradient.m: elementwise activations and derivatives.
- fmincg.m: conjugate-gradient optimiser.

2.7. Hyperparameters and Design Rationale

A hidden layer enables the network to capture non-linear relationships among question clusters while mapping them to the expert-defined impact scale. The hidden layer has 42 neurons, which coincides with the ordinal scale used for domain impact scores (1–42), but this is coincidental.

- **Lambda = 0.07** balances bias and variance for the small dataset; larger values over-smooth, smaller values overfit.
- **MaxIter = 10** is sufficient for convergence with fmincg on this objective; more iterations yield diminishing returns.
- Logistic outputs per domain fit the independence of targets and preserve the ability to calibrate or threshold each domain separately.

2.8. Training Procedure (Pseudocode)

```
procedure TRAIN(X, Y_scaled):
  initialise Theta1, Theta2 ~ small random
  options = { MaxIter: 10 }
  lambda = 0.07
  costFunction = @(t) nnCostFunction(t, input_size=4460, hidden_size=42,
num_labels=5, X, Y_scaled, lambda)
  theta_opt = fmincg(costFunction, [Theta1(:); Theta2(:)], options)
  reshape theta_opt -> Theta1, Theta2
  Y_hat_scaled = predict(Theta1, Theta2, X)
  return rescale_to_1_42(Y_hat_scaled)
```

 doi:10.20944/preprints202603.2068.v1

2.9. Computational Complexity and Runtime

Forward and backward passes are dominated by dense matrix multiplications of sizes (m × 4460) by (4460 × 42) and (m × 42) by (42 × 5). With m = 9, the runtime is negligible on commodity laptops; the memory footprint is dominated by the input matrix (approximately 9 × 4460 elements).

2.10. Data Preprocessing and Schema

- Inputs: nine CSVs X01.csv … X09.csv, each 223 rows by 20 columns, binary one-hot. These are flattened to a single 4460-length vector per project in column-major order.
- Labels: Y.csv, nine rows by five columns, integer scores on a 1–42 scale. Values are scaled to [0, 1] for training and rescaled for outputs.
- No imputation: only fully answered projects are included in this study.

3. Results

3.1. Overview of the Labelled Projects

The current dataset comprises nine citizen-science projects that completed the full MICS self-assessment and for which expert domain scores are available. Each project is represented by a 223 × 20 binary matrix of one-hot responses, which is flattened into a 4,460-dimensional vector. The label matrix contains five expert-assigned scores per project, corresponding to the Environment, Economy, Governance, Science, and Society domains, on an ordinal scale from 1 to 42.

Although the sample is small, the projects span a variety of designs, partnerships and intended outcomes. This diversity is essential for training Alquimics, because it exposes the network to different combinations of participation modes, data practices and impact pathways.

3.2. Training Behaviour and Internal Diagnostics

Because only nine projects were available, the primary focus of the present experiments is to verify that the model can be trained in a stable manner and produces sensible outputs. Training is performed using the conjugate-gradient optimiser (fmincg) for 10 iterations, with a regularisation strength of lambda = 0.07. During optimisation, the code monitors two quantities:

- The **loss function**, which combines the logistic cross-entropy terms for the five output nodes with the L2 weight penalty.
- The **RMSE** for each domain, computed on the training set and converted back to the original 1–42 scale of the expert scores.

In all runs inspected, the loss decreases monotonically over the first few iterations and then stabilises, indicating that the chosen learning setup is numerically well behaved. The RMSE values reported by the script are used as early indicators of whether particular hyperparameter settings or initialisations are clearly unsuitable. These diagnostics will later be complemented with cross-validated performance estimates (Sections 3.4 and 3.5).

3.3. Project-Level Predictions

After training on the full nine-project dataset, the network produces one five-dimensional output vector per project. Under LOOCV, each project's predicted scores are obtained from a model trained on the remaining eight projects, providing genuine out-of-sample estimates. Tables 1a and 1b present the expert-assigned (observed) scores and the corresponding LOOCV-predicted scores for all nine projects on the 1–42 scale.

 doi:10.20944/preprints202603.2068.v1

Table 1a. *Expert-assigned (observed) impact scores for all nine projects across five domains (1–42 scale).*

Project	Environment	Economy	Governance	Science	Society
P1	13	17	25	2	33
P2	25	2	6	20	37
P3	34	24	5	29	33
P4	18	36	11	34	21
P5	16	36	7	24	34
P6	26	17	9	4	12
P7	38	25	16	9	13
P8	18	24	22	1	34
P9	17	11	12	16	24

Table 1b. *LOOCV-predicted impact scores for all nine projects across five domains (1–42 scale). Each row is predicted by a model trained on the remaining eight projects.*

Project	Environment	Economy	Governance	Science	Society
P1	23	23	16	9	30
P2	17	26	15	8	32
P3	24	24	11	16	22
P4	26	17	10	15	29
P5	27	16	11	5	25
P6	28	27	11	17	26
P7	23	25	11	10	24
P8	15	13	19	8	33
P9	23	24	6	16	34

Several features of the predictions merit attention. **Range validity.** All predicted values lie within the valid 1–42 interval, confirming that the logistic output activations and label rescaling operate correctly across all nine held-out folds. **Regularisation pull.** Predicted scores cluster more tightly around the mid-range (roughly 10–28) than observed scores, which span nearly the full 1–42 range. This compression is most pronounced in the Economy domain, where extreme observed values (P2 = 2; P4 = 36; P5 = 36) are pulled substantially toward the centre by the L2 regularisation (λ = 0.07) and the limited training-set size (eight examples per fold). **Per-domain accuracy.** Governance predictions are closest to the observed scores across projects, while Economy and Science show the largest discrepancies, consistent with the domain-level RMSE values reported in Section 3.4. Predictions are written to "impact.csv" after rescaling, using the same 1–42 scale as expert labels.

3.4. Leave-One-Out Cross-Validation Results

Because only nine projects are available, LOOCV was adopted as the primary evaluation protocol. In each of the nine folds, the model was retrained from scratch on the eight remaining projects and applied to the held-out project; predictions were rescaled to the original 1–42 score range before computing errors.

Overall performance. Across all nine held-out predictions and five domains (45 data points), the LOOCV metrics on the 1–42 scale are: RMSE = 10, MAE = 9, R^2 = 0.06. The RMSE is close to the pooled standard deviation of the observed scores (SD = 11), indicating that the model's aggregate predictive power beyond a naïve mean prediction is modest. The low R^2 indicates that only a small fraction of the overall variance is explained, consistent with the severely constrained training regime (eight instances per fold for a 4,460-input model).

Domain-specific performance. Table 2 reports per-domain RMSE, MAE, and R^2 across the nine LOOCV folds. All values are on the original 1–42 impact scale.

 doi:10.20944/preprints202603.2068.v1

Table 2. *Per-domain LOOCV performance metrics (n = 9 folds, scores on 1–42 scale).*

Domain	RMSE	MAE	R^2
Environment	9	8	−0.2
Economy	14	11	−0.8
Governance	6	5	0.3
Science	12	10	−0.10
Society	9	8	0.02

Interpretation. Governance achieves the best predictive performance (RMSE = 6, R^2 = 0.3), suggesting that governance-related impact is more consistently encoded in the MICS questionnaire features. Economy and Science exhibit the weakest performance (RMSE > 12, negative R^2), indicating that the available features capture relatively less predictive signal for those domains or that expert scores for these domains vary more idiosyncratically across projects. The negative R^2 values for Environment, Economy, and Science indicate that, in those domains, the model performs worse than simply predicting the domain mean, an expected outcome when fitting a 4,460-dimensional model on eight training examples per fold. These results nonetheless serve as a transparent baseline: they quantify how well the current feature set and architecture generalise under the strictest data constraints, and they identify Governance as the most learnable domain with the present data. Full LOOCV will yield increasingly informative estimates as more labelled projects accrue.

3.5. Reproducibility Checklist

- Code available: Octave .m files included.
- Data schema documented: yes; nine input CSVs plus label matrix.
- Random seeds: set and logged in trainingMics.m.
- Environment: tested on Octave 10.3.
- Scripts: single-command run path documented; outputs written to impact.csv.
- Evaluation: LOOCV protocol and metrics specified.

4. Discussion

4.1. Key Findings and Model Design

Alquimics demonstrates that a compact multilayer perceptron can effectively map structured citizen-science project descriptors to domain-level impact assessments even within significant data constraints. The architecture employs five independent logistic output heads (one for each impact domain) rather than a single multi-class output layer. This design choice reflects both theoretical and practical considerations: it avoids imposing artificial trade-offs across conceptually distinct domains and aligns with downstream use cases in which IPAs must independently examine, compare, or threshold domain scores. The deliberately constrained model capacity (achieved through a modest hidden layer of 42 units), combined with strong regularisation techniques, prevents the network from overfitting to the small training set.

4.2. Methodological Contributions

The application of neural networks to citizen-science impact assessment marks a substantive methodological contribution. Traditional rule-based systems, while interpretable, cannot capture the complex, nonlinear relationships between project characteristics and multidimensional impact. Conversely, most machine-learning approaches to classification require substantially larger training datasets than were available here. Alquimics bridges this gap by demonstrating that a carefully designed neural network, constrained in capacity and regularised appropriately, can learn meaningful patterns from a limited but richly featured dataset. The use of approximately 200 input features represents the most comprehensive feature set assembled to date for citizen-science impact

 doi:10.20944/preprints202603.2068.v1

assessment, enabling the model to account for nuanced variations in project design, participation, data practices, and outcomes.

4.3. Critical Limitations: Data Constraints and Implications

The primary limitation is the small sample size: the training dataset comprised nine complete citizen-science projects. While this represents the most extensive collection of comprehensively documented projects available within MICS at the time of analysis, it raises substantive questions about model generalisation. With only nine instances, the model effectively learns patterns from a narrow slice of the global citizen-science landscape, potentially overlooking important project typologies, geographic variations, or implementation contexts not represented in the training set. This constraint necessitates considerable caution in applying Alquimics to projects substantially different from those in the training data.

A related concern is the potential presence of label noise: subtle inconsistencies or uncertainties in the expert-assigned impact scores used to train the model. Five domain scores were assigned for each project by domain experts via a rigorous evaluation protocol, yet human judgment inevitably contains subjectivity and measurement error. The impact of such noise on model learning deserves explicit discussion: in small-sample regimes, even modest label noise can substantially distort learned patterns. The regularisation strategies employed (weight decay, early stopping) provide some protection against this risk, but they cannot entirely eliminate it.

4.4. Temporal Scope and Long-Term Impact

A second major limitation concerns temporal scope. The MICS framework captures project characteristics and reported outcomes at a single time point or across limited timeframes, yet the true impact of citizen-science projects often unfolds over years or decades. For example, a project that stimulates new research directions, policy changes, or community engagement may show limited immediate impact but substantial long-term effects. Conversely, projects reporting high short-term metrics may not sustain engagement or real-world influence. Alquimics, trained on contemporaneous or near-term impact assessments, cannot reliably predict these longer-horizon effects.

4.5. Model Architecture and Validation

LOOCV results (Section 3.4) provide the honest out-of-sample baseline previously absent: overall RMSE = 10 on the 1–42 scale (cf. SD = 11), overall $R^2 = 0.06$, with Governance the most predictable domain (RMSE = 6, $R^2 = 0.3$) and Economy the least (RMSE = 14, $R^2 = -0.8$). These results confirm that the current model, trained on only eight examples per fold, does not yet generalise reliably, and should be interpreted as a transparent baseline rather than a validated predictive tool. Future work should expand the labelled dataset and systematically compare architectural choices across diverse projects and impact domains.

4.6. Practical Implications for MICS Users

Despite these limitations, Alquimics provides immediate practical value within the MICS ecosystem. The model generates domain-specific impact assessments and, when integrated with the rule-based recommendation engine, supplies actionable guidance to citizen-science project leaders. Users need not interpret Alquimics as a definitive or generalisable impact measure; rather, it functions as a learning tool, a system trained on documented projects that identifies patterns potentially relevant to new initiatives. This pragmatic framing acknowledges both the model's capabilities and its constraints, positioning it as one component of a broader assessment pipeline rather than a standalone solution.

 doi:10.20944/preprints202603.2068.v1

4.7. Broader Significance

This work demonstrates the feasibility and potential of machine learning for citizen-science impact assessment, even under data-limited conditions. It validates the value of comprehensive, structured data collection (reflected in MICS's 200+ features) and provides a methodological foundation for future, more robust systems. As citizen science continues expanding globally, the ability to quantify impact rigorously and comparatively across diverse projects becomes increasingly important for funding agencies, policy makers, and project communities. Alquimics represents an early, instructive step toward this goal, one that acknowledges both machine learning's promise for pattern discovery and the continued necessity of transparency, validation, and IPA engagement in assessment systems.

5. Limitations and Future Work

5.1. Limitations of the Present Study

The current analysis operates under three fundamental constraints that merit explicit acknowledgement. First, the training dataset comprises only nine citizen-science projects: the complete set available within MICS at the time of model development. While this represents an unusually rich and comprehensive collection (each with ~200 structured features), it remains limited for conventional machine-learning validation. Second, the impact scores assigned to these nine projects reflect expert judgment at a single time point, potentially missing long-term effects and introducing subjectivity or measurement error. Third, the projects in the training set may not be representative of the broader global citizen-science landscape, particularly with respect to geographic distribution, disciplinary focus, or implementation contexts. These constraints directly affect the generalisability of Alquimics predictions, particularly for novel projects that differ substantially from those in the training data.

5.2. Implications for Current Findings

Given these limitations, Alquimics should be interpreted not as a definitive, universally applicable impact-assessment system but as a domain-informed pattern-discovery tool trained on a specific collection of projects. The model's ability to identify relationships within the MICS dataset is genuine, but its performance on unseen projects, especially those with markedly different characteristics, remains uncertain. The framework is most defensible when applied to projects reasonably similar to those in the training set and when interpreted as one input to a broader assessment process alongside expert judgment and rule-based guidance.

5.3. Priority Directions for Future Work

Future research should pursue three complementary objectives: robustness validation, expanded training data, and methodological enhancement.

Robustness validation. The most immediate priority is systematic assessment of Alquimics's out-of-sample performance. LOOCV, in which the model is trained on eight projects and evaluated on the held-out ninth, repeated nine times, provides an unbiased estimate of how well the model generalises to unseen data. This approach maximises the use of limited training data while avoiding the over-optimistic performance estimates that can arise from training-set evaluation. LOOCV will yield domain-specific error metrics (root mean squared error, mean absolute error) that quantify prediction accuracy on the original impact scale. Additionally, sensitivity analyses should explore how model predictions respond to variations in hidden-layer width, regularisation strength, and the number of optimisation iterations. These ablation studies will reveal whether Alquimics's design is robust to reasonable hyperparameter variations or whether performance is brittle and dependent on specific tuning choices. Comparative evaluation against simpler baselines (linear regression, sparse models, regularised tree-based methods) will determine whether neural-network complexity is

 doi:10.20944/preprints202603.2068.v1

justified by improved predictions or whether signal in the feature set is already captured by interpretable, linear approaches.

Expanded training data. The fundamental limitation is the small sample size. Expanding the training dataset to 30–50 documented projects, a realistic target over a 2–3 year horizon, would substantially improve model stability, reduce overfitting risk, and enable meaningful train-validation-test splits. Critical to this expansion is intentional diversity: new projects should span different geographic regions, scientific disciplines, funding mechanisms, and community contexts to reduce the geographic and disciplinary biases likely present in the current nine-project sample. Equally important is standardisation of impact assessment; developing explicit rubrics for domain experts assigning scores would reduce label noise and improve consistency across projects.

Methodological enhancement. Three concrete improvements would strengthen the framework. First, integrating uncertainty quantification (either through Bayesian neural networks, Monte Carlo dropout at inference time, or ensemble methods) to provide confidence intervals around Alquimics predictions. This transparency is essential for IPAs' trust and decision-making. Second, implementing interpretability methods such as feature importance analysis or attention mechanisms to identify which project characteristics most strongly influence impact in each domain. Such insights would enrich both scientific understanding and practical guidance to users. Third, conducting longitudinal validation by tracking projects in the MICS database over time, comparing Alquimics's early-stage impact predictions against outcomes observed 12–36 months later. This would directly test whether the model captures drivers of lasting impact or merely correlates with short-term metrics.

5.4. Practical Roadmap

In the near term (6–12 months), LOOCV and ablation studies should be prioritised using an expanded project dataset and the archived code and data structures already prepared for this purpose. These analyses require no new data collection and will clarify the robustness and generalisability of the present model. In parallel, the documentation and labelling of new projects entering the MICS framework should be systematised to support incremental dataset expansion. In the medium term (1–2 years), as the training dataset reaches 25–40 projects, Alquimics should be re-trained and re-validated, incorporating uncertainty quantification and interpretability enhancements. In the longer term (2–3 years), sufficient labelled projects and temporal outcome data should be accumulated to conduct rigorous longitudinal validation and to explore semi-supervised or transfer-learning approaches that leverage unlabeled projects or related citizen-science datasets.

5.5. Alignment with MICS Ecosystem Needs

These research directions are not merely academic refinements; they directly serve the MICS user community. Project leaders and funders using MICS require not only impact scores but also transparency about model limitations and confidence. Expanding the training dataset and validating Alquimics across diverse project types will increase the relevance of recommendations to a global audience. Incorporating uncertainty quantification will support more nuanced decision-making. Interpretability analysis will help users understand *which* aspects of project design most strongly influence impact, enabling evidence-based improvement strategies. By pursuing these directions systematically, Alquimics will mature from a promising proof of concept into a trusted, widely applicable assessment tool.

6. Conclusions

A small, regularised neural network trained on structured self-assessment data can deliver useful, reproducible predictions of citizen-science impact across different domains. While the present evaluation is constrained by data availability, the results provide a credible baseline and a framework that scales cleanly as new labelled projects accrue.

 doi:10.20944/preprints202603.2068.v1

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Author Contributions: Conceptualisation: L.C.; Methodology: L.C., L.V.; Software: L.V., L.C., I.V.; Validation: L.C., L.V., I.V.; Formal analysis: L.C., L.V.; Investigation: L.C., L.V.; Data curation: L.C., L.V.; Writing – original draft: L.C., L.V.; Writing – review and editing: L.C., L.V., I.V.; Supervision: L.C.; Project administration: L.C. All authors have read and agreed to the published version of the manuscript.

Funding: The research described in this paper was funded by the European Commission via the MICS project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824711, and via the ProBleu project, which has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101113001 and from UK Research and Innovation under the UK government's Horizon Europe funding guarantee, grant number 10082336. The opinions expressed are those of the authors and are not necessarily those of the MICS or ProBleu partners, the European Commission, or the UK government.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Inputs are provided as X01.csv … X09.csv (each 223 × 20 binary), labels in Y.csv (nine rows × five domains), and the Octave implementation (mics.m, trainingMics.m, nnCostFunction.m, predict.m, sigmoid.m, sigmoidGradient.m, fmincg.m). Trained parameters are stored as Theta1.mat, Theta2.mat. A minimal run is:

```
% In Octave
>> mics % trains with MaxIter=10, writes impact.csv
```

The design matrix and scripts will be released in a public repository upon publication under the CC0 licence.

Acknowledgments: We thank the MICS consortium and the practitioners who completed the self-assessments that enabled this study. During preparation of this manuscript, the authors used an AI assistant (OpenAI ChatGPT; version November 2025) for language editing and formatting. The authors reviewed and edited all output and take full responsibility for the content.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Minimal Command-Line Recipes

```
# Train and write predictions
mics

# Change regularisation
edit trainingMics.m # set lambda and MaxIter, then rerun mics
```

Appendix B. Data Dictionary

XNN.csv: 223 rows by 20 columns, values in {0,1}. Row i corresponds to question i; columns represent categorical options.

Y.csv: nine rows by five columns. Columns are Environment, Economy, Governance, Science, and Society (integers 1–42).

Appendix C. Troubleshooting

If fmincg halts early, increase MaxIter to 20.

If costs explode, reduce lambda changes to smaller steps or re-initialise weights.

If predictions collapse to mid-range values, verify label scaling and gradient regularisation.

 doi:10.20944/preprints202603.2068.v1

References

Bonney, R., Shirk, J. L., Phillips, T. B., Wiggins, A., Ballard, H. L., Miller-Rushing, A. J., & Parrish, J. K. (2014). Next steps for citizen science. Science, 343(6178), 1436-1437.

Parkinson, S., Woods, S. M., Sprinks, J., & Ceccaroni, L. (2022). A Practical Approach to Assessing the Impact of Citizen Science towards the Sustainable Development Goals. Sustainability, 14(8), 4676.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), 160.

Sprinks, J., Woods, S. M., Parkinson, S., Wehn, U., Joyce, H., Ceccaroni, L., & Gharesifard, M. (2021). Coordinator perceptions when assessing the impact of citizen science towards sustainable development goals. Sustainability, 13(4), 2377.

Villena Román, J., Collada Pérez, S., Lana Serrano, S., & González Cristóbal, J. C. (2011). Hybrid approach combining machine learning and a rule-based expert system for text categorization. AAAI.

Xu, D., Shi, Y., Tsang, I. W., Ong, Y. S., Gong, C., & Shen, X. (2019). Survey on multi-output learning. IEEE transactions on neural networks and learning systems, 31(7), 2409-2429.