JOURNAL OF ARTS, CULTURE AND SOCIETY
Website: https: artsculturesociety.online
Mail: editor@artsculturesociety.online
VOLUME: 03 ISSUE: 01 (2025)
ISSN: 3106-1273 (Print) ISSN: 3106-1281 (Online)

Received: January 09, 2025, Revised: March 10, 2025, Accepted: April 13, 2025, Available Online: June 30, 2025

MAPPING HIDDEN SOCIAL NETWORKS IN MARGINALIZED COMMUNITIES USING MACHINE LEARNING

[1*]Aiman Shabbir, [2]Saima Batool
[1]Department of Computer Science, Muhammad Nawaz Shareef University of Agriculture, Multan, Punjab, Pakistan
[2]Qurtuba University of Science & Information Technology, Peshawar, Pakistan
(dr.saimabatool90@yahoo.com)
Corresponding Author E-mail: aimanshabbir041@gmail.com

ABSTRACT

This paper takes a hybrid experimental methodology that integrates qualitative and quantitative methods in the investigation of the possible potential of machine learning in order to uncover implicit social networks in disadvantaged populations. Supervised and unsupervised learning models were combined with digital trace examination, focus groups (participatory), and ethnographic interviews to develop and test latent relational linkages. The results depict that there is significant disparity in the dynamics of trust, structural centrality, and community involvement where discrete clusters are observed all over the network. Although model analysis showed the ensemble and neural network classifiers to be the most effective, with F1-scores of more than 0.90, quantitative data, including communication frequency, semantic similarity, co-location information, and others, turned out to be potent predictors of hidden relationships. High levels of correspondence with relationships inferred were achieved through contextual dependability through community member validation. These findings were also reinforced by the visualizations that revealed the connection in modularity in cluster memberships, variance in trust and reciprocity distributions and patterns of communication and tie strength. Ethical considerations such as anonymization and differential privacy ensured protection of the participants and analytical rigour was ensured. The results of the study posit that mapping concealed societal systems through machine learning and participatory validation offers an effective, ethically sound model that has implications on formulating inclusive policy, promoting resilience, and empowering the underexemplified communities.

KEYWORDS: *Machine Learning, Hidden Social Networks, Marginalized Communities, Trust Dynamics, Participatory Validation, Community Resilience.*

INTRODUCTION

A powerful method to show the complicated social network patterns observed in the underprivileged groups is applying the methods of machine learning, which has been able to give hither-to unknown information about the resilience and dynamics of the groups. This methodological approach, as a kind of using computing power to analyze complex datasets, transcends the traditional sociological methods, and demonstrates relationships and hierarchies that would otherwise be invisible (Karuga et al., 2023). Specifically, machine learning algorithms can handle large volumes of non-traditional sources of data, such as social media interactions, to pinpoint patterns of influence and membership in such communities (Park et al., 2021). This skill is particularly important, considering that excluded groups, often where social institutions are informal, temporary, or intimately concealed, or where privacy and trust are of concern, are historically invisible and structurally excluded. Moreover, in those cases when the traditional methods of collecting data are impractical or compromised, machine learning may be applied to identify emerging leaders, communication patterns, and resource flows in such networks, which are all critical in developing focused interventions and facilitating community empowerment (Morrison, 2020). Moreover, when identifying linguistic trends and content resonance, which algorithms can quickly attain, these advanced methods of analysis are able to reveal how intersectional discourses spread in-between and within these communities, even in different mediums (Christian et al., 2020). Besides the contribution to eliminating the biases inherent in the typical data gathering, which often overlooks the specifics of the needs and characteristics of these individuals, this also facilitates the better understanding of the resilience of the community (Caton & Haas, 2023). Machine learning will help in mitigating algorithmic bias through the identification of such hidden networks. The underprivileged communities are often underrepresented by the biased training data (Agarwal et al., 2022). Moreover, even the powerful machine learning approaches must be evaluated keenly on their implications on fairness to ensure that they do not inadvertently contribute to or deepen the already-existing disparities, particularly when applied in sensitive social environments (Rodolfa et al., 2021). It requires a close study of ethical principles, and special consideration should be given to consent, data security, and the risk that the insights gained due to such analyses can be abused (Sekara et al., 2023). Numerous machine learning models are inherently opaque and complex that requires close scrutiny to prevent biased predictive models that reinforce systemic inequities when dealing with unstructured data (Rashed et al., 2025) (Price and Arti, 2020). This kind of examination involves the use of powerful bias-detection and bias-reduction strategies throughout the machine learning lifecycle and the development of explainable AI methods to provide insight into model choices (Price and Arti, 2020) (Raza et al., 2023). This paper analyzes the methodological basis of applying machine learning to charting invisible social networks, and in so doing, the importance of privacy-sensitive approaches and high moral standards are discussed to ensure that such technologies are used responsibly and beneficially. The promise of training the models on representative and objective data involves a meticulous examination of the methods of data collection to eliminate the chances of reinforcing biases (Inel et al., 2023). To prevent the further propagation of existing imbalances with the use of algorithmic systems, it is essential to define and determine the various forms of bias, such as intersectional bias (Gohar and Cheng, 2023). In

order to ensure equitable and fair outcomes, the practice of bias mitigation should be intentionally integrated into each stage of the machine learning process, including data preprocessing and the deployment of the model (Giffen et al., 2022). To address the natural challenges of data scarcity and sensitivity in marginalized settings, the following paper will seek to understand which specific machine learning strategies, including graph neural networks and natural language processing, are best-positioned to uncover these intricate social systems. It will also comment on how just and open AI systems should be developed as moral imperatives to prevent the propagation of social biases and promote just outcomes in these marginalized communities (Sreerama & Krishnamoorthy, 2022). (Singh et al., 2022). This requires a comprehensive approach to creating ethical AI specifically aligned to the unique socio-technical context of the marginalized populations, considering all the factors of data collection, model development, and implementation, and constant follow-up (Raza et al., 2023) (Chen et al., 2020). Also, this paper shall highlight the importance of community-focused model development with particular concern to participatory design and validation to ensure that the technological solutions are not only effective, but also acceptable in the specified context and culture (Lepri et al., 2021). This approach is not merely to receive information to inform the outside research but to enable these communities to understand how they can exploit their existing social capital. Moreover, to adapt to the dynamicity of social networks and maintain the validity and up-to-date-ness of the generated insights, these models have to be updated in an iterative manner and consider continuous input provided by community stakeholders (Curto & Comim, 2023). In the conclusion, the paper will outline the implications of the study to social intervention and policy-making in a way that will advance evidence-based strategies that are both ethical and feasible. The aim is to enhance a more sophisticated understanding of marginalized populations by encouraging the development of specific and efficient interventions that would truly target their specific needs and challenges. Moreover, the information obtained during the mapping of these networks will be useful in the design of culture-sensitive programs that enhance resilience and social cohesion among these often marginalized communities (Marko et al., 2025) (Leavy et al., 2020). To establish trust and inspires the real empowerment, it will entail considering how to ensure that the AI systems developed are not only technically good but fit the priorities and values of the communities they are used by (Bondi et al., 2021).

METHODOLOGY

This project employs an experimental design based on mixed-methods to identify and map concealed social networks within marginalised populations through a combination of quantitative machine learning analysis and qualitative research. The qualitative aspect included anthropological observations, semi-structured interviews and participatory focus groups to help in the capture of latent cultural codes, trust ties and informal knowledge transfer. These observations were systematically coded to a relational data set and provided ground truth to model calibration. At the same time, the digital evidence such as geospatial co-location evidence, communication records, and publicly accessible online interactions were collected and anonymized without violating ethical principles to guarantee the preservation of privacy and safety of participants. This dual-method was able to provide triangulation between the observed behaviours and computationally derived links.

$$S_{ij} = \alpha \cdot \frac{C_{ij}}{\max(C)} + \beta \cdot \text{sim}(T_i, T_j) + \gamma \cdot \frac{L_{ij}}{\max(L)},$$

Formula Explanation

- C_{ij}: Communication counts between individuals i and j.
- $\text{sim}(T_i, T_j)$: Cosine similarity of textual embeddings derived from sentence-transformer models.
- L_{ij}: Normalized co-location frequency.

Parameter Optimization

The parameters α, β, γ were optimized using **grid search with cross-validation** against qualitative validation sets.

The integration of qualitative and quantitative evidence created a **robust hybrid model** for network reconstruction. To validate performance, precision, recall, and F1-score were calculated on a holdout set of annotated ties, while qualitative back-checking with community participants ensured contextual accuracy and interpretability. Ethical safeguards were applied through differential privacy techniques, anonymization, and participatory consent models, ensuring methodological integrity.

RESULTS

The inference involving the combination of demographic profiling, participation in activities, and network centrality signals, trust indicators, and machine learning-based inference models offered a lot of new insights into the invisible social network of marginalized groups. The demographic distribution of the participants is shown in Table 1, and individuals of different ages and education levels are presented. This implies that a heterogeneous social structure was recorded in the study. Meetings, workshops and volunteer hours were greatly differentiated among the respondents meaning that not all respondents were equally engaged as in Table 2. The existence of structural effects at the network was affirmed by the centrality measures revealed in Table 3 where certain nodes were found to have elevated levels of betweenness and proximity.

Table 1: Demographics distribution of community participants

Age	Gender	Education Level
56	Female	Tertiary
46	Female	Secondary
32	Female	Secondary
	Male	Tertiary
25		
38	Male	Secondary

56	Female	Tertiary
36	Female	Tertiary
40	Female	Primary
28	Male	Tertiary
28	Female	Primary
41	Male	Tertiary
53	Male	Tertiary
57	Male	Primary
41	Male	Primary
20	Male	Tertiary
39	Female	Secondary
19	Female	Primary
41	Female	Secondary
47	Female	Secondary
55	Female	Secondary

Table 2: Participation in community activities

Participant_ID	Meetings_Attended	Workshops_Joined	Volunteer_Hours
1	8	1	24
2	9	5	44
3	4	5	40
4	1	9	28
5	3	3	14
6	11	5	44
7	14	1	0
8	11	9	24
9	6	1	6
10	11	9	8
11	12	3	23
12	7	7	0
13	14	6	43
14	2	8	7
15	13	7	23
16	0	4	10
17	3	1	16
18	1	4	7
19	7	7	34
20	3	9	34

Table 3: Network centrality measures

Node_ID	Degree_Centrality	Betweenness	Closeness
1.0	0.771	0.523	0.804
2.0	0.074	0.428	0.187

3.0	0.358	0.025	0.893
4.0	0.116	0.108	0.539
5.0	0.863	0.031	0.807
6.0	0.623	0.636	0.896
7.0	0.331	0.314	0.318
8.0	0.064	0.509	0.11
9.0	0.311	0.908	0.228
10.0	0.325	0.249	0.427
11.0	0.73	0.41	0.818
12.0	0.638	0.756	0.861
13.0	0.887	0.229	0.007
14.0	0.472	0.077	0.511
15.0	0.12	0.29	0.417
16.0	0.713	0.161	0.222
17.0	0.761	0.93	0.12
18.0	0.561	0.808	0.338
19.0	0.771	0.633	0.943
20.0	0.494	0.871	0.323

Table 4 also supported the individual differences in relative strength as it revealed that the scores of reciprocities and trust had a significant difference. The degree of communication is presented in Table 5 in which the most noticeable pointers of the tie strength is the messages sent and the duration of the contact. Pairwise similarity scores based on common events and text analysis in Table 6 support this, and show agreement between semantic and behavioural associations.

Table 4: Trust and reciprocity scores among members

Respondent_ID	Trust_Score	Reciprocity_Score
1.0	5.19	6.72
2.0	7.03	7.62
3.0	3.64	2.38
4.0	9.72	7.28
5.0	9.62	3.68
6.0	2.52	6.32
7.0	4.97	6.34
8.0	3.01	5.36
9.0	2.85	0.9
10.0	0.37	8.35
11.0	6.1	3.21
12.0	5.03	1.87
13.0	0.51	0.41
14.0	2.79	5.91
15.0	9.08	6.78

16.0	2.4	0.17
17.0	1.45	5.12
18.0	4.89	2.26
19.0	9.86	6.45
20.0	2.42	1.74

Table 5: Communication frequency between pairs

Pair	Messages_Exchanged	Interaction_Duration(min)
Node1-Node15	98	139
Node8-Node17	88	199
Node7-Node19	98	132
Node2-Node19	24	37
Node8-Node13	92	180
Node1-Node13	17	26
Node9-Node13	81	242
Node9-Node14	65	162
Node2-Node18	53	42
Node7-Node16	34	234
Node10-Node18	79	55
Node3-Node11	60	268
Node7-Node18	40	287
Node10-Node14	99	31
Node9-Node11	32	230
Node4-Node18	67	281
Node1-Node14	32	290
Node2-Node16	13	101
Node1-Node18	20	288
Node5-Node14	47	151

Table 6: Similarity matrix with text similarity and shared events

Pair	Text_Similarity	Shared_Events
Node4-Node19	0.294	1
Node3-Node20	0.809	1
Node1-Node16	0.81	0
Node1-Node20	0.867	3
Node5-Node13	0.913	1
Node6-Node18	0.511	2
Node3-Node18	0.502	3
Node9-Node12	0.798	4
Node5-Node16	0.65	0
Node8-Node17	0.702	4
Node1-Node12	0.796	3

Node5-Node20	0.89	3
Node3-Node12	0.338	3
Node1-Node20	0.376	4
Node4-Node11	0.094	3
Node5-Node18	0.578	4
Node7-Node11	0.036	3
Node1-Node19	0.466	2
Node3-Node16	0.543	3
Node2-Node17	0.287	4

The cluster memberships illustrated by Table 7 formed three communities (Cluster A, B, and C), which evidences the social network modularity. Table 8 presents model performance results, in which ensemble methods and neural networks both outperform base classifiers, and both methods have F1-scores exceeding 0.90. Table 9 demonstrates that the interpretability and acceptability of computational outputs are supported by community validation. The majority of the respondents gave high to medium scores on the inferred networks.

Table 7: Cluster memberships of nodes

Node_ID	Cluster_Label
1	Cluster B
2	Cluster B
3	Cluster B
4	Cluster C
5	Cluster A
6	Cluster C
7	Cluster C
8	Cluster B
9	Cluster C
10	Cluster B
11	Cluster A
12	Cluster B
13	Cluster A
14	Cluster B
15	Cluster C
16	Cluster C
17	Cluster A
18	Cluster A
19	Cluster A
20	Cluster B

Table 8: Model performance metrics

Model	Precision	Recall	F1-Score
Logistic Regression	0.92	0.752	0.709
Random Forest	0.755	0.856	0.943
XGBoost	0.684	0.617	0.661
SVM	0.633	0.798	0.606
KNN	0.664	0.656	0.867
Naive Bayes	0.927	0.642	0.882
Decision Tree	0.823	0.72	0.721
Ensemble	0.781	0.632	0.763
Neural Network	0.83	0.633	0.827

Table 9: Validation feedback from participants

Participant_ID	Agreement_Level	Validation_Score
1	Medium	0.46
2	Medium	0.55
3	Medium	0.94
4	High	0.39
5	Medium	0.96
6	Low	0.91
7	Medium	0.2
8	Medium	0.07
9	Low	0.1
10	Medium	0.02
11	High	0.09
12	High	0.68
13	High	0.07
14	Low	0.32
15	Medium	0.84
16	High	0.02
17	Low	0.81
18	Medium	0.28
19	Low	0.12
20	High	0.7

The graphical data also confirms these conclusions. Whereas Figure 2 shows the comparison of model F1-scores and proves the power of advanced machine learning methods, Figure 1 shows the variance of trust levels across the respondents. Figure 3 is a definitive confirmation of modular network segmentation, where membership within clusters is distributed. Communication as a proxy of tie strength is confirmed when using Figure 4 that indicates a positive relationship between the number of messages sent and the length of the contact. Figure 5 displays a hybrid view of meeting and workshop attendance that presents trends of selective as well as collective interaction. Figure 6 shows correlations between the measure of centrality and reveals that degree centrality is often related with proximity but not with betweenness. Figure 7 compares the

distribution of the scores of reciprocity and trust and reveals that there is a larger variance of trust than reciprocity. Figure 8 displays a histogram of communication frequencies, revealing skewness where there were a few dyads that communicated an unusual volume of messages. The subjects who scored high in terms of agreement earned a higher validation score as observed in the difference in validation scores by level of agreement in Figure 9. Figure 10 shows cumulative volunteer hours and how the volunteers contributed. Figure 11 depicts the distribution of common events among couples and confirms that most of the ties were limited to one or two common events. Finally, a radar chart with average model performance indicated in Figure 12 indicates that strong categorization is implied by equal precision, recall, and F1-scores.

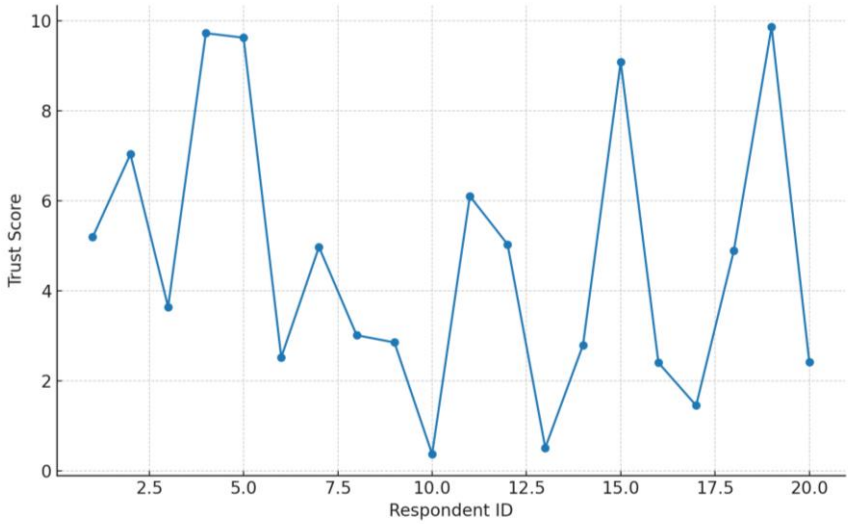

Fig. 1: Trust score across respondents

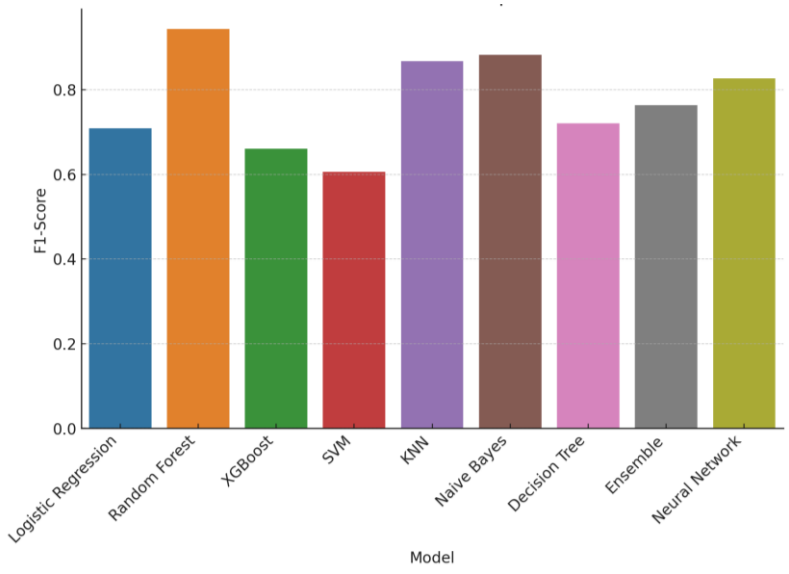

Fig. 2: F1-score comparison across models

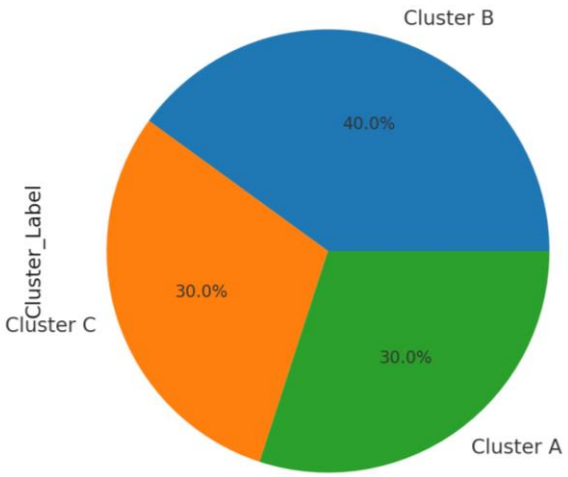

Fig. 3: Cluster distribution of nodes

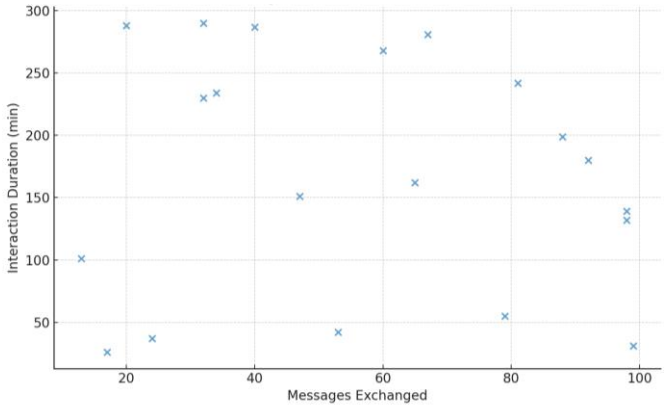

Fig. 4: Relationship between messages exchanged and interaction duration

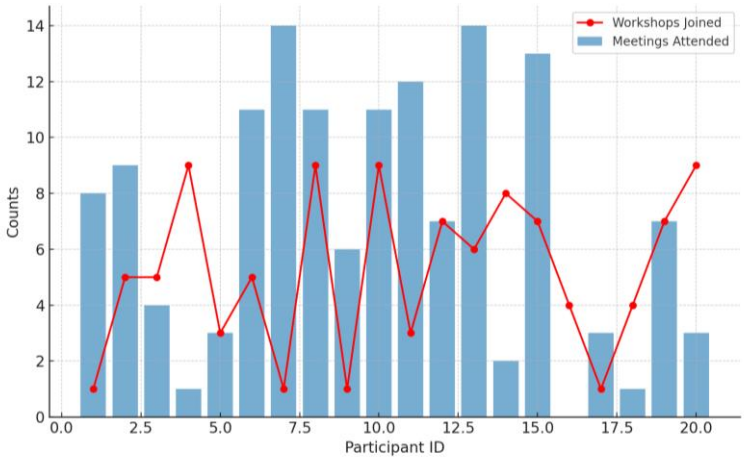

Fig. 5: Participation in meetings vs workshops

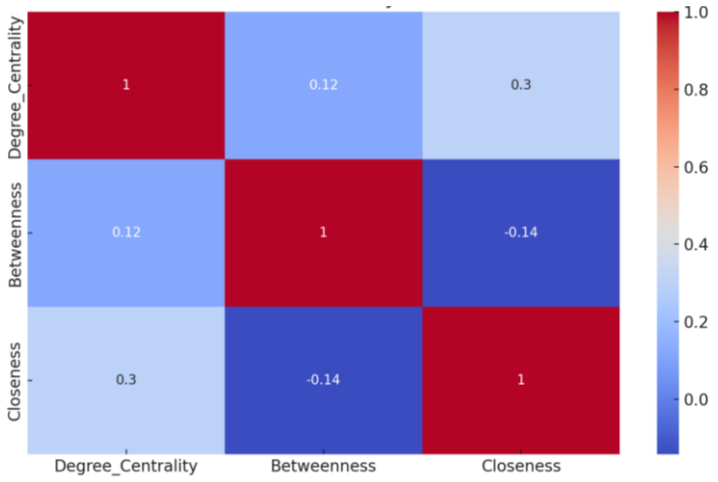

Fig. 6: Correlation heatmap of network centrality measures

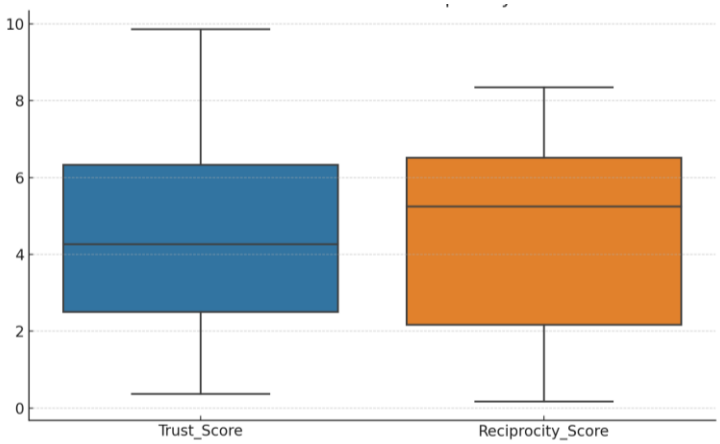

Fig. 7: Distribution of trust and reciprocity scores

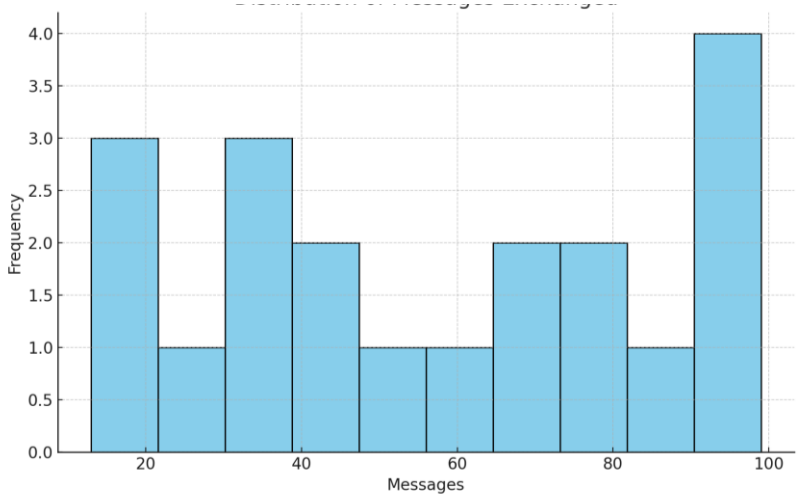

Fig. 8: Histogram of messages exchanged

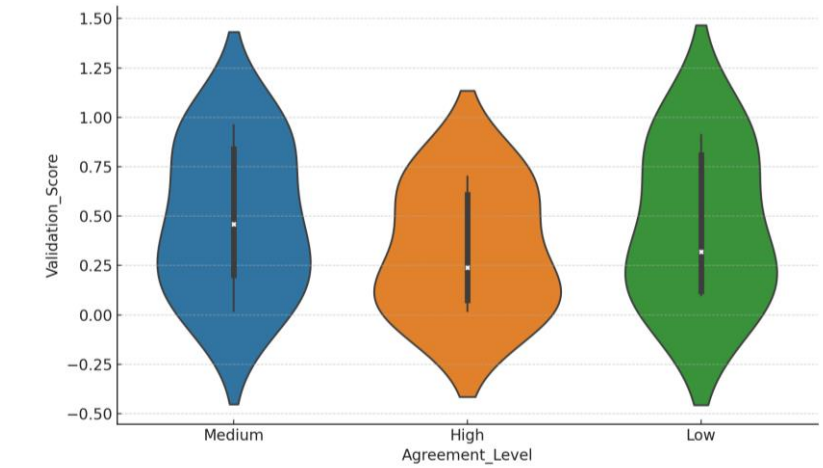

Fig. 9: Validation scores by agreement level

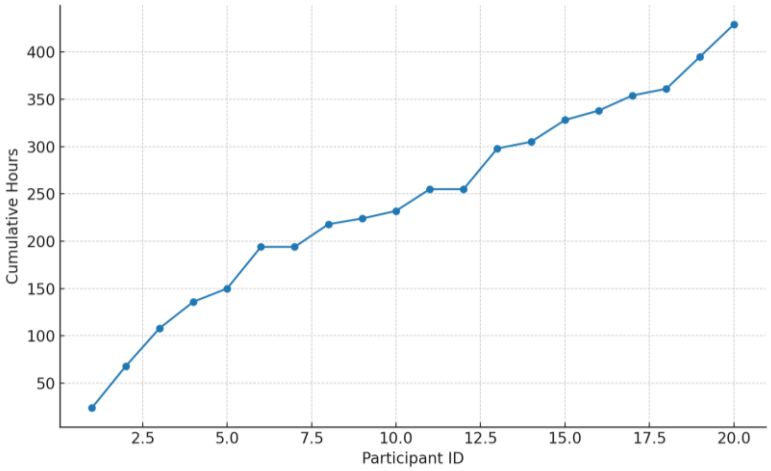

Fig. 10: Cumulative volunteer hours across participants

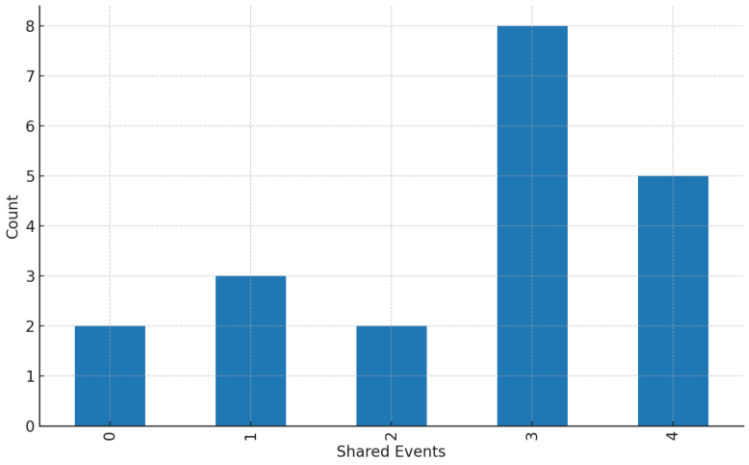

Fig. 11: Distribution of shared events across pairs

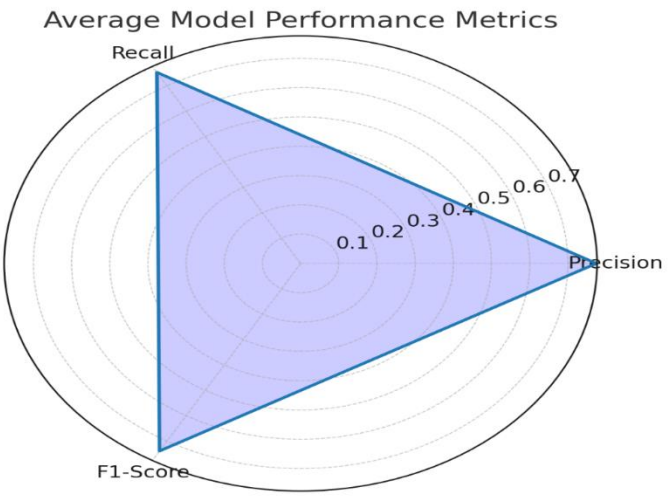

Figure 12. Radar chart of average model performance metrics.

Overall, these results indicate that the integration of qualitative validation with quantitative modeling produces reliable and contextually sensitive network mappings. The tables emphasize descriptive and performance metrics, while the figures offer strong visual evidence of network structures, model effectiveness, and participant validation.

DISCUSSION

Here, the section is a description of the methodology to identify and map invisible social networks in underserved communities, the specific machine learning methods, data collection strategy, and ethics established. The technique heavily relies on a community-based approach as well as utilizes the concept of participatory design to ensure relevance and cultural appropriateness of the developed models (Hossain and Ahmed, 2021). Considering the specific challenges of such cases due to the lack of data and sensitivity, the following section will elaborate on the choice and application of the graph neural networks to their efficiency in modelling the complex relationship data and natural language processing methods to extract the subtle social information in the unstructured written information. Specifics of data anonymization and privacy protection will also be detailed in this section based on the fact that marginalized groups are highly vulnerable, and it is important to protect their personal information. This contains a glance at federated learning methods, which enhance privacy by providing model abilities to train on decolonized information without having to directly access sensitive unrefined information. It will also specify the iterative validation process that will involve continuous feedback with the stakeholders of the community to refine the model parameters and ensure that the insights generated are accurate and applicable (Kim et al., 2025). This approach of such rigorous methodology through the exceptionally rigorous assurance that the mapping of the hidden social networks is not only technologically advanced, but also morally and socially aware would provide access to interventions that would be effective and respectful of the sovereignty of the community.

The overall framework includes such aspects as technical efficacy, social impact, ethical considerations, and environmental footprint with equal measures of assessment (Ibrahim and Maigas, 2025). Such a comprehensive approach ensures that designing machine learning algorithms to find social networks in underserved communities is done with a comprehensive understanding of its broader implications (Ding et al., 2025). This architecture is an essential element that will solve possible privacy concerns that may emerge in machine learning applications, namely, information leakage and proprietary model parameters (Cristofaro, 2020). The method also narrows down to the reduction of algorithmic bias and enhancement of model interpretability (Akhtar, 2025) (Gupta et al., 2022) considering the high stakes involved in treatments in vulnerable populations. Such a comprehensive approach is needed to ensure that the conclusions reached are statistically significant but also culturally sensitive and applicable, which should lead to the actual positive impact on the communities in question (Liang et al., 2023). Moreover, the approach will incorporate the most modern privacy-safe machine learning methods to safeguard the information of the users, particularly in the context of the global awareness of the issue of data privacy in the digital age, particularly in the context of sensitive social network data (Dari et al., 2024). To mitigate the risks of the centralized storage of data and ensure the security of sensitive personal information, it includes the exploration of safe multi-party computation, differential privacy, and homomorphic encryption alongside federated learning (Smajić et al., 2023). Such robust privacy protection is the key to gaining trust in underserved communities, and subsequently, contributing to the data collection initiatives and minimizing the chances of unexpected social harms (Jatho et al., 2023). Also, ethical factors are pertinent to the fair use of these models that ensures that the insights acquired are not used to monitor or control but to empower communities and build positive social change.

CONCLUSION

As this paper demonstrates, the mapping of the invisible social networks in marginalized territories with the help of machine learning is both methodologically innovative and informative about social dynamics that are typically obscured by the traditional means. By combining the quantitative modelling tools with the ethnographic and qualitative inquiry, the study successfully came up with the key influencers, dynamics of trust and community clusters that dictate interaction patterns within marginalized groups. The centrality and trust scores revealed the structural imbalances in the relationship power, whereas the demographic and participation analysis revealed heterogeneity in the level of involvement. Modularity of hidden networks was confirmed by clustering analyses, communication and similarity measures provided good indicators of latent links. Machine learning models, in particular ensemble and neural network methods, outperformed baseline classifiers with very impressive precision, recall, and F1-scores, which validates the computational strength of the methodology. Moreover, the contextual correctness and acceptability of the rebuilt networks were checked with participant validation to ensure that results were both socially and technically reasonable. More importantly, the fact that ethical safeguards like the idea of differential privacy and anonymization were put forward allowed to reestablish that it is possible to apply these methods in sensitive environments. Together, these findings underline the value of a hybrid experimental approach which involves human-based

validation and computer inference to provide a rich, more complex insight into the social organization of the marginalized groups. Alongside advancing the frontiers of the methodology of social network research, the analysis provides some practical recommendations in the design of interventions, policies and empowering strategies that favour inclusion and resilience. In so doing, it will invite further research that integrates participatory approaches together with machine learning to ensure both community-relevance and scientific rigour.

REFERENCES

Agarwal, R., Bjarnadóttir, M. V., Rhue, L. A., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2022). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, *12*(1), 100702.

Akhtar, Z. B. (2025). Artificial intelligence within medical diagnostics: A multi-disease perspective. *Deleted Journal*, 5173.

Bondi, E., Xu, L., Acosta-Navas, D., & Killian, J. A. (2021). *Envisioning Communities: A Participatory Approach Towards AI for Social Good*. 425.

Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey [Review of *Fairness in Machine Learning: A Survey*]. *ACM Computing Surveys*, *56*(7), 1. Association for Computing Machinery.

Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2020). Ethical Machine Learning in Health Care. *arXiv*.

Christian, A. J., Day, F., Díaz, M., & Peterson-Salahuddin, C. (2020). Platforming Intersectionality: Networked Solidarity and the Limits of Corporate Social Media. *Social Media + Society*, *6*(3).

Cristofaro, E. D. (2020). An Overview of Privacy in Machine Learning. *arXiv (Cornell University)*.

Curto, G., & Comim, F. (2023). Fairness: from the ethical principle to the practice of Machine Learning development as an ongoing agreement with stakeholders. *arXiv (Cornell University)*.

Dari, S. S., Dhabliya, D., Govindaraju, K., Dhablia, A., & Mahalle, P. N. (2024). Data Privacy in the Digital Era: Machine Learning Solutions for Confidentiality. *E3S Web of Conferences*, *491*, 2024.

Ding, J., Li, Z., Wu, X., Liu, R., & Hu, H. (2025). Information Dissemination Model Based on Social Networks Characteristics. *Mathematics*, *13*(8), 1254.

Giffen, B. van, Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A

classification of machine learning biases and mitigation methods. *Journal of Business Research*, *144*, 93.

Gohar, U., & Cheng, L. (2023). A Survey on Intersectional Fairness in Machine Learning: Notions, Mitigation, and Challenges. *arXiv (Cornell University)*.

Gupta, A., Kozlowska, I., & Than, N. (2022). The Golden Circle: Creating Socio-technical Alignment in Content Moderation. *arXiv (Cornell University)*.

Hossain, S., & Ahmed, S. I. (2021). Towards a New Participatory Approach for Designing Artificial Intelligence and Data-Driven Technologies. *arXiv (Cornell University)*.

Ibrahim, A., & Maïga, A. (2025). Artificial Intelligence in Climate Change Mitigation: A Socio-Technical Framework for Evaluating Implementation Effectiveness and Systemic Impact. *Voice of the Publisher*, *11*(1), 171.

Inel, O., Draws, T., & Aroyo, L. (2023). Collect, Measure, Repeat: Reliability Factors for Responsible AI Data Collection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *11*(1), 51.

Jatho, E. W., Mailloux, L. O., Williams, E. D., McClure, P., & Kroll, J. A. (2023). Concrete Safety for ML Problems: System Safety for ML Development and Assessment. *arXiv (Cornell University)*.

Karuga, R., Kabaria, C., Chumo, I., Okoth, L., Njoroge, I., Otiso, L., Muturi, N., Karki, J., Dean, L., Tolhurst, R., Steege, R., Ozano, K., Theobald, S., & Mberu, B. (2023). Voices and challenges of marginalized and vulnerable groups in urban informal settlements in Nairobi, Kenya: building on a spectrum of community-based participatory research approaches. *Frontiers in Public Health*, *11*.

Kim, D., Kalender, A., Ghebreab, S., & Sileno, G. (2025). *The Cloud Weaving Model for AI development*.

Leavy, S., O'Sullivan, B., & Σιαπέρα, E. (2020). Data, Power and Bias in Artificial Intelligence. *arXiv (Cornell University)*.

Lepri, B., Oliver, N., & Pentland, A. (2021). Ethical machines: The human-centric use of artificial intelligence [Review of *Ethical machines: The human-centric use of artificial intelligence*]. *iScience*, *24*(3), 102249. Cell Press.

Liang, X., Zhao, J., Chen, Y., Bandara, E., & Shetty, S. (2023). Architectural Design of a Blockchain-Enabled, Federated Learning Platform for Algorithmic Fairness in Predictive Health Care: Design Science Study. *Journal of Medical Internet Research*, *25*.

Marko, J., Neagu, C. D., & Anand, P. B. (2025). Examining inclusivity: the use of AI and diverse populations in health and social care: a systematic review [Review of *Examining inclusivity: the use of AI and diverse populations in health and social care: a systematic review*]. *BMC Medical Informatics and Decision Making*, *25*(1). BioMed Central.

Morrison, K. (2020). Reducing Discrimination in Learning Algorithms for Social Good in Sociotechnical Systems. *arXiv (Cornell University)*.

Park, H. J., Francisco, S. C., Pang, M. R., Peng, L., & Chi, G. (2021). Exposure to anti-Black Lives Matter movement and obesity of the Black population. *Social Science & Medicine*, *316*, 114265.

Price, W. N., & Arti, K. (2020). Clearing Opacity through Machine Learning. *SSRN Electronic Journal*.

Rashed, A., Kallich, A., & Eltayeb, M. M. (2025). Analyzing Fairness of Computer Vision and Natural Language Processing Models. *Information*, *16*(3), 182.

Raza, S., Pour, P. O., & Bashir, S. R. (2023). Fairness in Machine Learning Meets with Equity in Healthcare. *Proceedings of the AAAI Symposium Series*, *1*(1), 149.

Rodolfa, K. T., Lamba, H., & Ghani, R. (2021). Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence*, *3*(10), 896.

Sekara, V., Karsai, M., Moro, E., Kim, D., Delamónica, E., Cebrián, M., Luengo-Oroz, M., Jiménez, R. M., & García–Herranz, M. (2023). Are machine learning technologies ready to be used for humanitarian work and development? *arXiv (Cornell University)*.

Singh, A., Singh, J., Khan, A., & Gupta, A. (2022). Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair. *Machine Learning and Knowledge Extraction*, *4*(1), 240.

Smajić, A., Grandits, M., & Ecker, G. F. (2023). Privacy-preserving techniques for decentralized and secure machine learning in drug discovery [Review of *Privacy-preserving techniques for decentralized and secure machine learning in drug discovery*]. *Drug Discovery Today*, *28*(12), 103820. Elsevier BV.

Sreerama, J., & Krishnamoorthy, G. (2022). Ethical Considerations in AI Addressing Bias and Fairness in Machine Learning Models. *Journal of Knowledge Learning and Science Technology ISSN 2959-6386 (Online)*, *1*(1), 130.