

A Safety-First Machine Learning Framework for Precision Public Health: Segmenting Vaccine Hesitancy Among Iranian Adults

Mehrdad Askarian

Shiraz University of Medical Sciences

Shamila Bayati

Amirkabir University of Technology

Hakim Rajabi

Amirkabir University of Technology

Ardalan Askarian

University of Saskatchewan

Nahid Hatam

Shiraz University of Medical Sciences

Nazanin Ayareh

nazanin.ayareh@gmail.com

Shiraz University of Medical Sciences

Research Article

Keywords: Vaccine Hesitancy, Machine Learning, Precision Public Health, Psychological Reactance, Explainable AI

Posted Date: August 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10306536/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Public Health on October 9th, 2026.
See the published version at <https://doi.org/10.1186/s12889-026-29763-2>.

Abstract

Background

Traditional one-size-fits-all public health communication often fails to account for individual psychological barriers, potentially triggering psychological reactance and backfire effects. To address this, a Precision Public Health approach is required to categorize populations into tailored communication paths based on their unique trust profiles.

Methods

Utilizing data of a cross-sectional study of 457 Iranian adults, we developed a computational framework to map objective demographic and geographic signatures onto three psychometric segments: Accepting, Ambivalent, and Resistant. The framework utilized a theory-grounded rule-based mapping (HBM and 3C models) to define target classes, followed by a Safety-First machine learning pipeline (Ensembled Logistic Regression and HistGradientBoosting). To address selection bias inherent in our highly educated, healthcare-heavy cohort (46.4% medical professionals), an internal fold-based preprocessing pipeline was strictly enforced. To prioritize ethical safety, a confidence threshold ($\tau = 0.53$) was implemented as a Safety Valve to re-route statistically ambiguous classifications into a human-centered, dialogue-based intervention path.

Results

While the model achieved a modest accuracy of 47.28%, it represents a 42% relative improvement over the random baseline (33.33%). Explainable AI (SHAP) analysis revealed that Province of Residence and Age were the primary determinants of segment membership. The τ threshold re-routed approximately 20% of uncertain cases to the Ambivalent segment, significantly reducing critical errors and ensuring the framework defaults to engagement under conditions of uncertainty.

Conclusion

This study provides a validated descriptive blueprint for risk-stratified health communication. By shifting the paradigm from raw accuracy to Safety-First triage, the framework offers a scalable method for health authorities to identify and address specific psychological barriers within the Iranian population, enhancing future pandemic preparedness through empathetic, data-informed intervention.

Background

Vaccines have served as the primary defense in the global effort to control and end the COVID-19 pandemic, preventing severe illness and mortality on a population scale.(1, 2) However, vaccine

hesitancy has severely restricted public engagement, which is crucial to the success of mass immunization efforts.(3) These public health outcomes are driven by mechanisms such as institutional trust and acceptability; voluntary adherence to guidelines is more successfully attained when trust in health authorities is high.(4)

Communication efforts have been hampered in the Iranian setting by differences in trust, particularly the difference between national health systems and international medical organizations.(5, 6) Traditional public health strategies have historically relied on a one-size-fits-all approach, utilizing generic mass-communication campaigns.(5, 7) The failure of conventional, generic messaging poses a serious risk to public safety in addition to being a logistical inefficiency.(4) Recent research indicates that these approaches frequently overlook personal psychological obstacles.(8) Psychological reactance, a condition in which people feel threatened by their autonomy and take a stance that is opposite to the one being promoted, can be triggered by sending the same high-pressure message to a heterogeneous population.(4, 9) This has the unintended consequence of pushing hesitant or undecided people toward resistance rather than compliance, which worsens social polarization and reduces vaccination rates overall.(5)

A Precision Public Health strategy is needed to reduce this risk.(10) This calls for a need-based grouping approach that customizes therapies to fit particular demographic and psychological characteristics. Age and sex are the main factors influencing vaccine side effects, according to recent studies that successfully used machine learning (ML) to predict clinical reactogenicity.(11, 12) Although these tools facilitate tailored clinical insights, there is still a significant gap in the literature; there are currently no validated computational frameworks that can map various signatures to unique individual communication needs.

In order to solve these issues, this study has two main goals. To create a machine learning pipeline that uses demographic and geographic features for trust profiles to divide Iranian adults into actionable policy groups, and to validate a Safety-First triage model by applying a probability threshold. While the descriptive analysis provided by Askarian et al. (2025), builds upon that foundation by transforming static behavioral predictors into a dynamic Precision Public Health triage framework. By mapping objective demographic signatures onto latent trust profiles, we provide a computational architecture capable of individual-level communication routing. Furthermore, we address the "mediocre" model fit inherent in behavioral research by implementing a mathematical "Reject Option". This ensures that algorithmic uncertainty defaults to human-centered dialogue, mitigating the risk of triggering psychological reactance through misaligned automated messaging.

Methods

1. Study Design and Data Source

Data were gathered from a sample of 457 Iranian individuals between September and November 2024 using a cross-sectional study approach published earlier.(5) An anonymous online survey was used for the study, and it was shared on popular digital channels like Instagram, WhatsApp, and Telegram. The Health Belief Model (HBM) (13) and the 3C Model of Vaccine Hesitancy(14) (Confidence, Complacency, and Convenience) served as the theoretical foundation for the survey instrument, guaranteeing that the variables recorded were consistent with accepted behavioral theories concerning public health compliance.(5)

2. Feature Engineering and Segmentation

The main function of the framework is to convert psychometric data into segments of practical policy. Three composite indices, Acceptance, Institutional Trust, and Doubt, were developed based on the 3C and HBM frameworks to prevent unjustified labeling. Rule-Based Mapping was used to divide participants into three different classes. The 33rd and 66th percentiles (quantiles) of the composite score distribution were used to establish boundaries. Doubt was measured by the frequency of neutral replies, Institutional Trust concentrated on trust in national health systems, and Acceptance included items about intent to vaccinate and perceived efficacy.

Class 1 (Accepting): High trust and acceptance scores (> 66th percentile).

Class 2 (Ambivalent): Moderate trust and acceptance scores (33rd–66th percentile)

Class 3 (Resistant): Low trust and high clinical doubt (< 33rd percentile).

3. Data Preprocessing and Normalization

A multi-stage preprocessing was constructed in Python to guarantee the machine learning pipeline's stability. A custom normalization method was used to standardize raw Persian text replies to accommodate unnecessary whitespace, different forms of the letters Ye and Ke, and zero-width non-joiners (ZWNJ). Additionally, responses were translated to numerical indices (1–5) using a 5-point Likert scale, which ranges from Strongly Disagree to Strongly Agree. We used a hybrid imputation approach to maintain sample size and data quality. The median was used for missing numerical values, and the mode for categorical variables, which was encapsulated strictly within each cross-validation fold. This ensured that the statistical distribution of the test set did not affect the training phase.

4. Machine Learning Architecture

We used the Scikit-Learn toolkit to create a modular machine learning pipeline.(15) To manage the simultaneous processing of numerical and category information, the architecture used a Column Transformer.

Three main algorithms were assessed and ensembled by the framework; Table 1 summarizes their initial performance metrics and compares them to a random baseline.

- 1. Logistic Regression: Selected for its linear interpretability and high performance in healthcare settings.(16)
- 2. HistGradientBoosting (HGB): Utilized to capture non-linear relationships and interactions within the demographic data.(17)
- 3. Voting Classifier: A soft-voting ensemble was implemented to aggregate the predicted probabilities of the base models, providing the final confidence score for each classification.(18)

Table 1
Model Performance

Model Architecture	Mean Accuracy	Mean Macro F1-Score	Standard Deviation (σ)
Logistic Regression (Tuned)	0.4604	0.4566	0.032
Soft Voting Ensemble	0.4678	0.4479	0.038
HistGradientBoosting	0.4505	0.4323	0.041
<i>Random Guess Baseline</i>	<i>0.3333</i>	<i>0.3333</i>	-

To guarantee generalizability, the model's performance was verified using stratified 5-fold cross-validation. We used the `class_weight='balanced'` parameter for all models to address class imbalance within the Resistant section.

5. Confidence Threshold (τ)

The application of a Safety Valve triage rule, which acts as an ethical fallback for algorithmic decision-making, is a key novelty of this approach. Even in cases where the underlying data is unclear, standard machine learning algorithms force every individual into a category. To reduce the dangers of misclassification, our method adds a confidence threshold (τ). The implementation of the τ threshold is grounded in the foundational 'Reject Option' framework, which establishes the mathematical trade-off between classification error and the decision to defer in cases of high uncertainty.(19, 20) The algorithm automatically overrides the initial prediction and places the person in Class 2 if the model's maximum probability score for a particular individual is less than τ (optimal at 0.53 in our study). The mathematical optimization of τ was explicitly bounded to minimize “Critical Errors”, defined as misclassifying a Resistant individual (Class 3) as Accepting (Class 1), thereby ensuring the pipeline operates as a true safety fallback.

Results

1. Demographic and Class Profile

The study population (N = 457) represented a diverse cross-section of the Iranian adult population. The median age of participants was 48 years (IQR: 38–57), with a balanced gender distribution (53.8% female). A significant portion of the sample consisted of highly educated individuals, with 63% holding a master's degree or higher, and 46.4% identifying as healthcare professionals. Residency was geographically distributed over several locations, with provinces like East Azerbaijan, Isfahan, and Ardabil having the greatest participation rates.

Initially, to assess the separability of the behavioral signatures, the population was evaluated across a granular 5-tier spectrum (Classes 1 to 5). Multi-class ROC curves were generated for this preliminary 5-class distribution, demonstrating substantial discriminative ability at certain intervals, particularly for Classes 4 and 5 (AUC = 0.735 and 0.680, respectively). However, to translate these findings into actionable and robust public health interventions, and to establish the ground-truth targets for the demographic mapping task, this continuum was aggregated into three broader policy segments using the 33rd and 66th percentiles: Class 1 (Accepting): n = 146 (~32%), Class 2 (Ambivalent): n = 178 (~39%), and Class 3 (Resistant): n = 133 (~29%) (Figure 1), demonstrating that the features successfully captured unique behavioral signatures, particularly for the extreme Resistant segment (Class 3).

2. Model Performance and Baseline Comparison

The classification framework, using a tuned Logistic Regression pipeline with internal imputation, achieved an overall Accuracy of 47.28% and a Macro F1-score of 45.97% at the optimal threshold. The model's achievement of a 14-percentage-point margin over the random guess baseline (33.33%) confirms that objective demographic variables offer a statistically detectable signal regarding segment membership. By identifying high-risk or high-need groups that would otherwise go unnoticed under "one-size-fits-all" tactics, even models with modest predictive power can greatly optimize population-level resource allocation.⁽²¹⁾ Figure 2 shows how the probability threshold and model metrics relate to one another.

The cumulative confusion matrix analysis (Figure 3) confirmed that the model prioritizes clinical safety. Misclassifications between nearby segments (Class 3 to Class 2) were mostly Safe Errors. Importantly, the model rarely misclassified Resistant people as Accepting (Class 1), achieving excellent specificity for Class 3.

3. Determinants of Adherence

We used SHAP (Shapley Additive Explanations) to quantify feature influence in order to determine the objective demographic anchors for subjective trust profiles (Figure 4). By determining each variable's distinct contribution to the model's output, this method offers a mathematically rigorous approach to feature importance.⁽²²⁾ Three main factors that function as predictive signatures for triage were found by this analysis.

- 1. Geography: The best predictor was living in a particular province. There were noticeable changes in the baseline likelihood of adherence among people in regions such as East Azerbaijan and Isfahan. SHAP values indicated that residency in these specific provinces acted as a primary anchor for the model’s classification logic, outperforming generalized demographic variables like gender or income.
- 2. Experience: The impact of a family history of COVID-19 deaths was significant. Personal loss served as a reality check, according to SHAP values, which moved people away from theoretical uncertainty and toward a greater likelihood of following vaccination schedules.
- 3. Demographics: Younger populations (18–34) and specific job sectors (non-healthcare roles) were strong filters for Class 2 and Class 3 classifications.

4. Threshold Analysis

The main method for controlling uncertainty was the application of the confidence threshold (τ). Table 2 provides a summary of the mathematical trade-offs between algorithmic coverage and policy-safe accuracy. As demonstrated:

At low τ : The chance of critical errors was higher, but coverage was high (almost the whole population received a particular 1, 2, or 3 label).

At $\tau = 0.53$ (Optimal) The framework re-routed around 20% of the most ambiguous cases to Class 2. While this reduces the coverage of the autonomous segments, it significantly increases the clinical reliability of the remaining interventions. The resulting shift in population distribution, the re-routing of uncertainty to the Dialogue segment, is illustrated in Figure 5, while the overall efficiency of the search range is summarized in Figure 6.

Table 2 Threshold Optimization Analysis

Threshold (τ)	Safe Accuracy	Macro F1-Score	Policy Implication
0.3	0.4579	0.4543	High coverage; high risk of mislabeling resistant cases.
0.38	0.4628	0.4597	Optimal for Macro F1: Balanced class sensitivity.
0.53	0.4728	0.4401	Optimal for Safety: Highest safe accuracy; routes ~20% to Dialogue.
0.7	0.4282	0.3194	Excessive caution; degrades personalized utility.

Discussion

Traditional public health messaging has proven inadequate for addressing the heterogeneous psychological barriers that drive vaccine hesitancy, often exacerbating resistance through psychological

reactance.(4, 7) The purpose of this study was to ascertain if geographic and demographic variables might be used as trustworthy anchors to divide the population into discrete groups with varying communication needs. The main conclusion is that vaccine hesitation and resistance are consistently linked to geographic and demographic groups within the Iranian population rather than being randomly distributed. The Province of Residence was the most significant predictor of a trust profile, according to our SHAP research. There were noticeable changes in the baseline likelihood of segment membership among people in regions such as East Azerbaijan and Isfahan. This research highlights the local nature of trust in Iran, indicating that social, linguistic, and cultural factors greatly influence the demand for health communication.(23)

This geographic variance suggests that a centralized, one-size-fits-all message designed in a capital center may be perceived as culturally distant. Such messaging can cause psychological reactance in places where there is less national trust, when people take a defiant stand because they believe the communication threatens their authority. (4) The Ministry of Health can prioritize decentralized efforts that use local influencers to disseminate health information by using our system to detect regional trust signatures.

The Class 2 (Ambivalent) segment should be viewed as a population caught in a "clash of trusts". By identifying these individuals via geographic signatures, our framework allows for decentralized interventions that utilize local influencers to bridge the transparency gap and mitigate the risk of psychological reactance.

The model determined that age is a crucial segmentation criterion. The tendency for younger populations to fall into high-support or resistant segments is not only a matter of opinion but is likely rooted in physical experience. According to clinical evidence, young people experience more systemic side effects, including fever, chills, and headaches, and reactogenicity after immunization than elderly. (11, 12) Public health messages that ignore these frequent side effects in favor of concentrating only on the vaccine's safety cause a reality gap for younger people, which increases reluctance. Our proposed framework shifts communication from general safety claims to an empathetic validation of clinical experiences by mapping younger age signatures to higher-need segments. In these cases, healthcare providers ought to offer clear recommendations on managing short-term side effects to lessen anxiety.

Finding demographic factors that differentiate the Ambivalent group (Class 2) through information-seeking behavior was a major accomplishment of the computational process. Class 2 people actively seek out validated information to clear up their clinical doubt, whereas Class 3 people frequently show complete disengagement from institutional health sources. When allocating resources, this distinction is crucial. By using demographic filters to identify this population, health workers can give them the expert access and technical depth they need before they become resistant.

While an accuracy of 47.28% is numerically modest, it is significantly robust given the aleatoric uncertainty inherent in human behavioral data. Beating the baseline by this margin demonstrates that the model has detected valid behavioral signatures, which is important in the field of human behavioral

research because multicausality and irreducible noise impose a low upper bound on predictability. (24, 25)

However, in public health, the cost of a False Positive is high.(26) Our τ threshold serves as an ethical safety valve to lessen this. Treating an unsure person as Ambivalent (Class 2) and directing them toward conversation is far safer from a triage point of view than mislabeling them as Resistant (Class 3) and triggering a backfire effect. This default to dialogue rule ensures that algorithmic uncertainty results in human-centered engagement rather than high-pressure automated messaging. However, we acknowledge that sending low-confidence forecasts to Class 2 inevitably reduces this segment's accuracy. A combination of people with statistically unclear trust signatures and those who were genuinely cautious make up the resulting Ambivalent group. In order to optimize the clinical safety of the triage system, this decrease in segment purity is a trade-off.

Our model offers health professionals a useful manual that goes beyond theoretical analysis. Resources can be prioritized based on the unique psychological obstacles of each group by dividing the population into segments. We suggest an organized communication strategy, which is described in the Policy Action Matrix (Table 3), to convert these computational results into proactive public health management.

Table 3
Policy Action Matrix

Predicted Segment	Confidence Level	Strategy	Targeted Communication Tactic
Class 1 (Accepting)	High ($P \geq \tau$)	Facilitation	Automated SMS with direct scheduling links and where/when info.
Class 2 (Ambivalent)	High ($P \geq \tau$)	Reassurance	Interactive FAQ bots and links to peer-reviewed Iranian clinical data.
Class 3 (Resistant)	High ($P \geq \tau$)	Trust Building	Engagement via non-governmental community leaders; autonomy-respecting tone.
Any Class	Low ($P < \tau$)	Safety Default	Automatic route to Dialogue (Class 2): Connect with a medical professional.

Limitations

Although the framework is strong, there are a few limitations that must be acknowledged. First, the cross-sectional methodology of the study only offers a temporal picture of public trust, which is dynamic. Additionally, the sample's demographic makeup also exhibits selection bias because it overrepresents healthcare professionals and highly educated people, which might not represent varied Iranian population. Specifically, the high density of healthcare professionals (46.4%) and highly educated participants (63%) may result in an 'insider bias,' where the sample's baseline trust levels are higher than those of the broader, more heterogeneous Iranian population. Lastly, the small sample size prohibits the

use of more data-intensive deep learning architectures, even though it is adequate for the machine learning techniques used here.

Conclusion

According to our study, vaccine hesitancy and resistance are not random occurrences but are associated with specific demographic and geographic groups in the Iranian population. By switching from generic mass-messaging to a computational triage framework, we have demonstrated the reliability of mapping objective signatures onto latent trust profiles, particularly age and residency. Rather than absolute prediction accuracy, the main benefit of this technique is its capacity to identify the demographic factors that influence institutional skepticism and public health compliance. An important ethical development is the application of the Safety-First paradigm through the τ threshold. To guarantee that algorithmic uncertainty leads to a default to the conversation rule, the framework prioritizes preventing psychological reactance over maximizing autonomous coverage. In the end, this framework offers a road map for regionalized, decentralized communication tactics that tackle the clinical reality of age-dependent reactogenicity and the localized nature of trust. It also serves as a model basis for the upcoming generation of Precision Public Health tools, which will combine behavioral ethics and data science. By mathematically optimizing the trade-off between autonomous classification and the 'Reject Option,' the framework provides a rigorous method for managing the ethical risks of AI-driven public health interventions.

Abbreviations

3C: Confidence, Complacency, and Convenience Model

AUC: Area Under the Curve

FAQ: Frequently Asked Questions

HBM: Health Belief Model

HGB: HistGradientBoosting

IQR: Interquartile Range

ML: Machine Learning

ROC: Receiver Operating Characteristic

SHAP: SHapley Additive exPlanations

SMS: Short Message Service

τ (tau) : Confidence Threshold

Declarations

Ethics approval and consent to participate

The original survey was reviewed and approved by the Ethics Committee of Shiraz University of Medical Sciences (Approval ID: IR.SUMS.REC.1403.430). All participants provided electronic informed consent prior to completing the questionnaire. The present secondary analysis was conducted in full compliance with the ethical principles of the Declaration of Helsinki.

Consent for publication

Not applicable.

Clinical trial number

Not applicable.

Availability of data and materials

The dataset supporting the conclusions of this article is available from the corresponding author upon reasonable request, subject to ethical and institutional data sharing policies.

Competing interests

The authors declare that they have no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

NA conceived the study idea based on the available data and designed the overall framework. NA also served as the primary point of communication for the team and drafted the original manuscript. SB and HR developed the computational code, implemented the machine learning pipeline, and performed the formal analysis. AA provided technical guidance and critical revisions on the computational and methodological aspects of the manuscript. MA provided the dataset, supervised the research process, and made substantial contributions to the interpretation of the results. NH contributed to the revision of the manuscript and provided critical feedback on the public health implications. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank all the participants who generously took the time to complete the survey.

References

1. Cheon S, Methiyothin T, Ahn I. Analysis of COVID-19 vaccine adverse event using language model and unsupervised machine learning. *PLoS ONE*. 2023;18(2):e0282119.
2. Watson OJ, Barnsley G, Toor J, Hogan AB, Winskill P, Ghani AC. Global impact of the first year of COVID-19 vaccination: a mathematical modelling study. *Lancet Infect Dis*. 2022;22(9):1293–302.
3. Cerda AA, García LY. Hesitation and refusal factors in individuals' decision-making processes regarding a coronavirus disease 2019 vaccination. *Front public health*. 2021;9:626852.
4. Hosseinabadi ZA, Brantnell A. Sweden's public health response to COVID-19: a qualitative study building on a realist approach. *BMC Health Serv Res*. 2025;25(1):1393.
5. Askarian M, Ahmadkhani A, Mousavi S, Muhajarine N, Sadeghi A, Zarei S, et al. Adherence to national and international vaccine information sources and future pandemic preparedness among Iranian adults in the post-COVID-19 era. *BMC Public Health*. 2025;25(1):2859.
6. Zarocostas J. How to fight an infodemic. *lancet*. 2020;395(10225):676.
7. Berlemann M, Haustein E. Right and yet wrong: a spatio-temporal evaluation of Germany's COVID-19 containment policy. 2020.
8. Coleman S, Slater MD, Wright P, Wright O, Skardon L, Hayes G. Pandemic lifeworlds: A segmentation analysis of public responsiveness to official communication about Covid-19 in England. *PLoS ONE*. 2024;19(1):e0296049.
9. Brehm SS, Brehm JW. Psychological reactance: A theory of freedom and control. Academic; 2013.
10. Khoury MJ, Galea S. Will precision medicine improve population health? *JAMA*. 2016;316(13).
11. Eterafi M, Fouladi N, Amanzadeh M, Mohammadnia A, Safarzadeh E. A data-driven machine learning framework to predict side effects of AstraZeneca and sinopharm COVID-19 vaccines. *Sci Rep*. 2025;15(1):40687.
12. Jamshidi E, Asgary A, Kharrazi AY, Tavakoli N, Zali A, Mehrazhi M, et al. Personalized predictions of adverse side effects of the COVID-19 vaccines. *Heliyon*. 2023;9(1):e12753.
13. Rosenstock IM. Historical origins of the health belief model. *Health Educ Monogr*. 1974;2(4):328–35.
14. MacDonald NE. Vaccine hesitancy: Definition, scope and determinants. *Vaccine*. 2015;33(34):4161–4.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–30.
16. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206–15.

17. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W et al. Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30.
18. Dietterich TG, editor. Ensemble methods in machine learning. International workshop on multiple classifier systems; 2000: Springer.
19. Chow C. On optimum recognition error and reject tradeoff. *IEEE Trans Inf Theory*. 2003;16(1):41–6.
20. Kompa B, Snoek J, Beam AL. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digit Med*. 2021;4(1):4.
21. Kleinberg J, Ludwig J, Mullainathan S, Obermeyer Z. Prediction policy problems. *Am Econ Rev*. 2015;105(5):491–5.
22. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30.
23. Damari B, Sharifi V, Asgardoost MH, Hajebi A. Community Action Package in Iran's Comprehensive Mental and Social Health Services (the SERAJ Program). *Iran J Psychiatry*. 2021;16(1):76.
24. Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect Psychol Sci*. 2017;12(6):1100–22.
25. Salganik MJ, Lundberg I, Kindel AT, Ahearn CE, Al-Ghoneim K, Almaatouq A et al. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*. 2020;117(15):8398 – 403.
26. Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science*. 2019;363(6429):810–2.

Figures

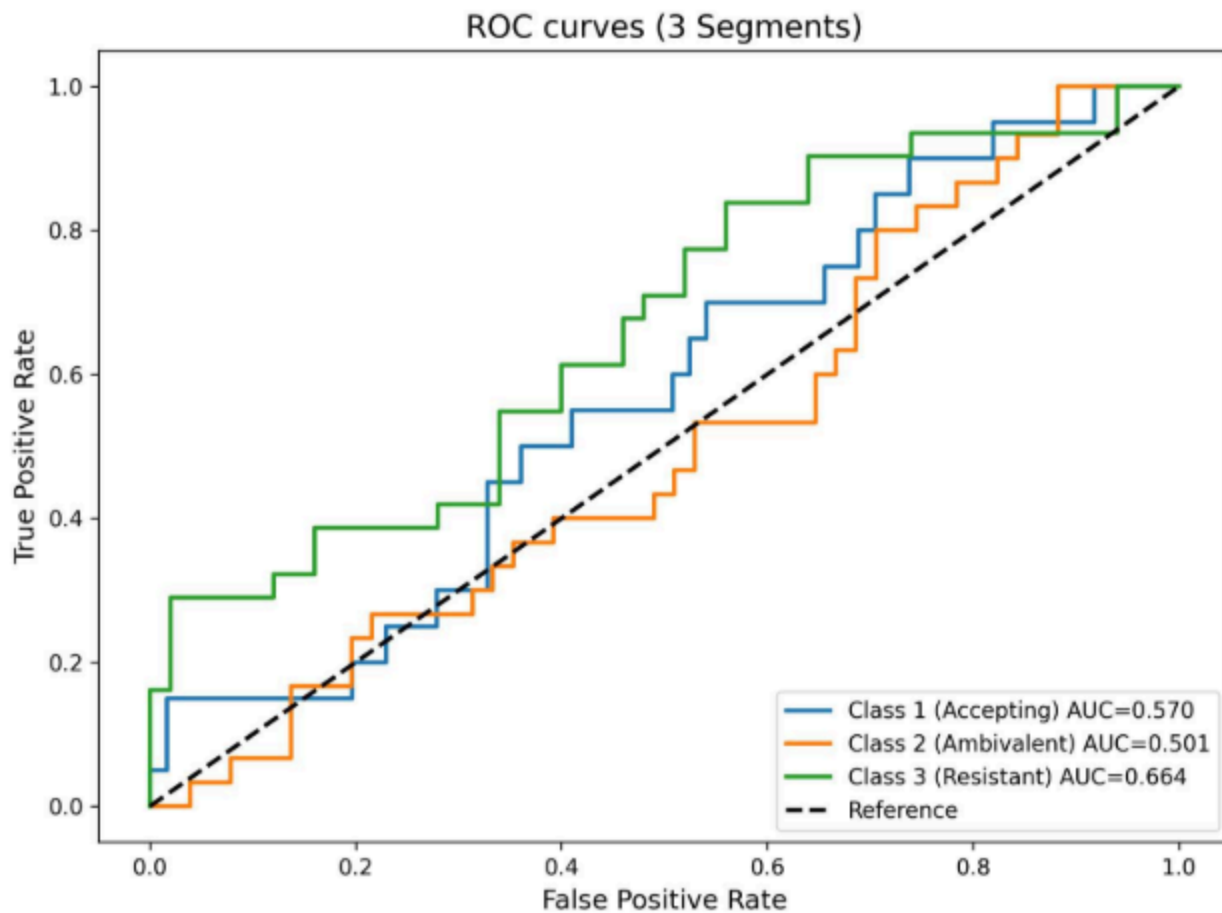


Figure 1

Receiver Operating Characteristic (ROC) Curves.

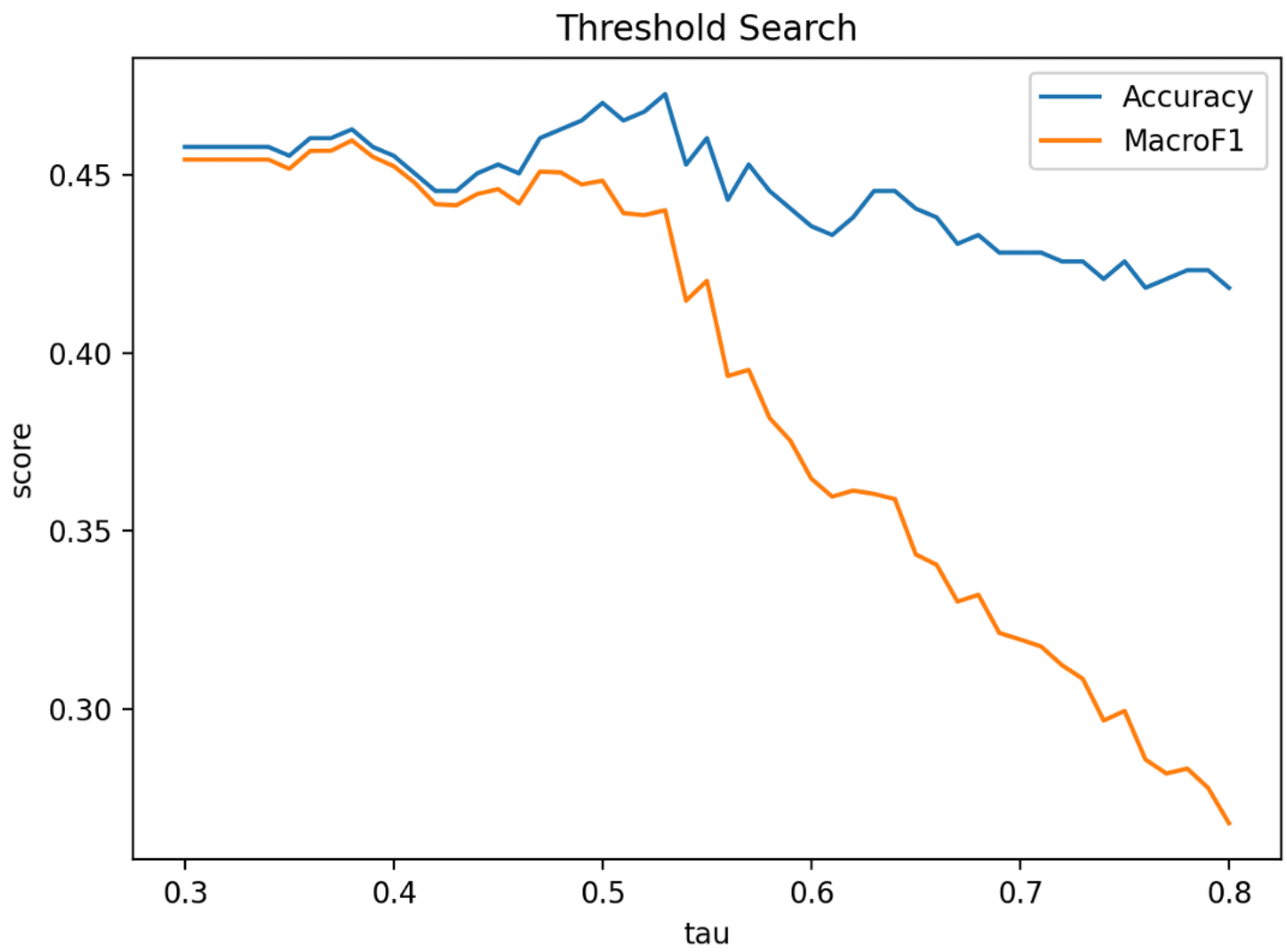


Figure 2

Optimization of the Confidence Threshold (τ) for the Safety-First Mechanism.

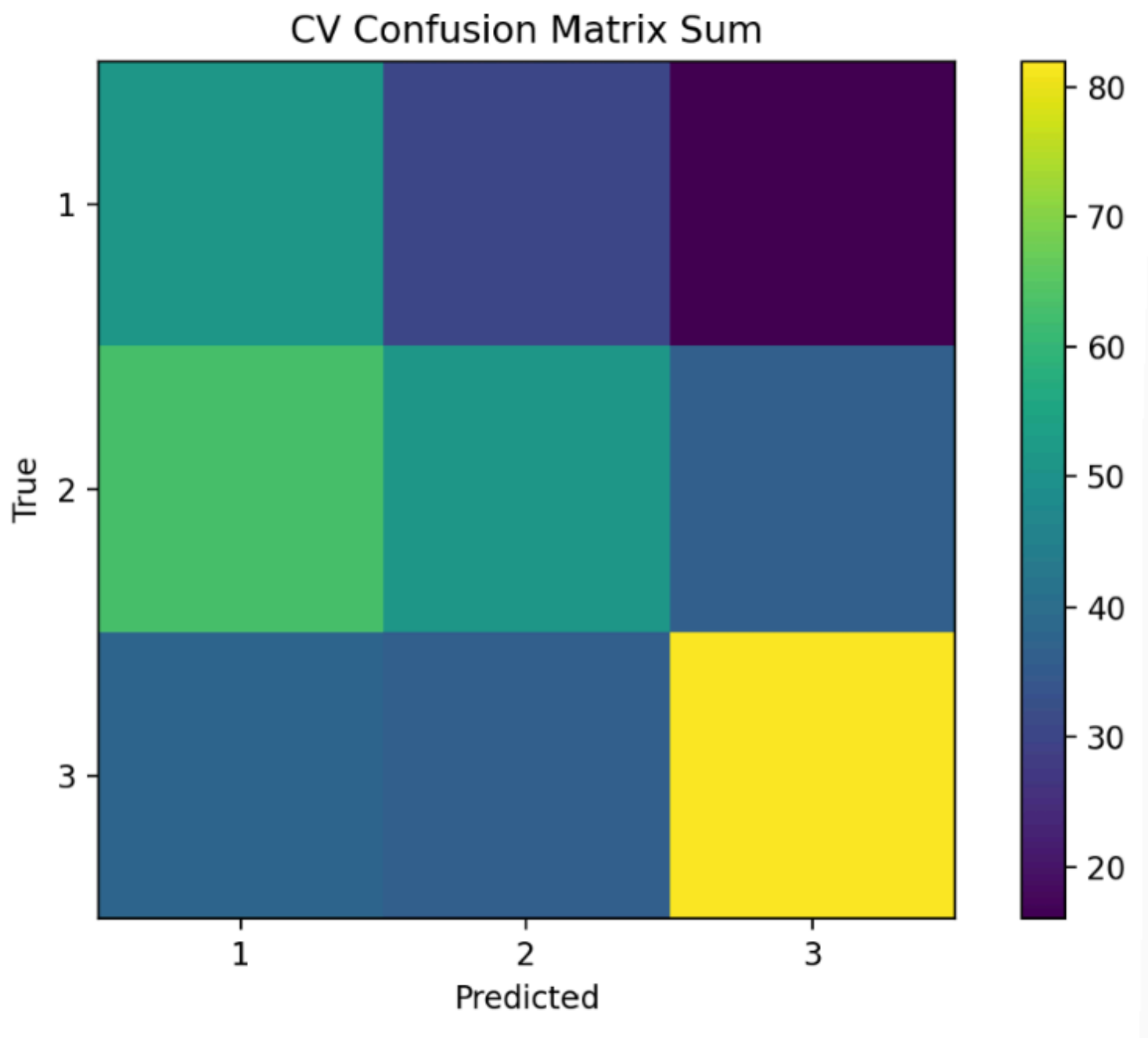


Figure 3

Cumulative Confusion Matrix across Stratified 5-Fold Cross-Validation.

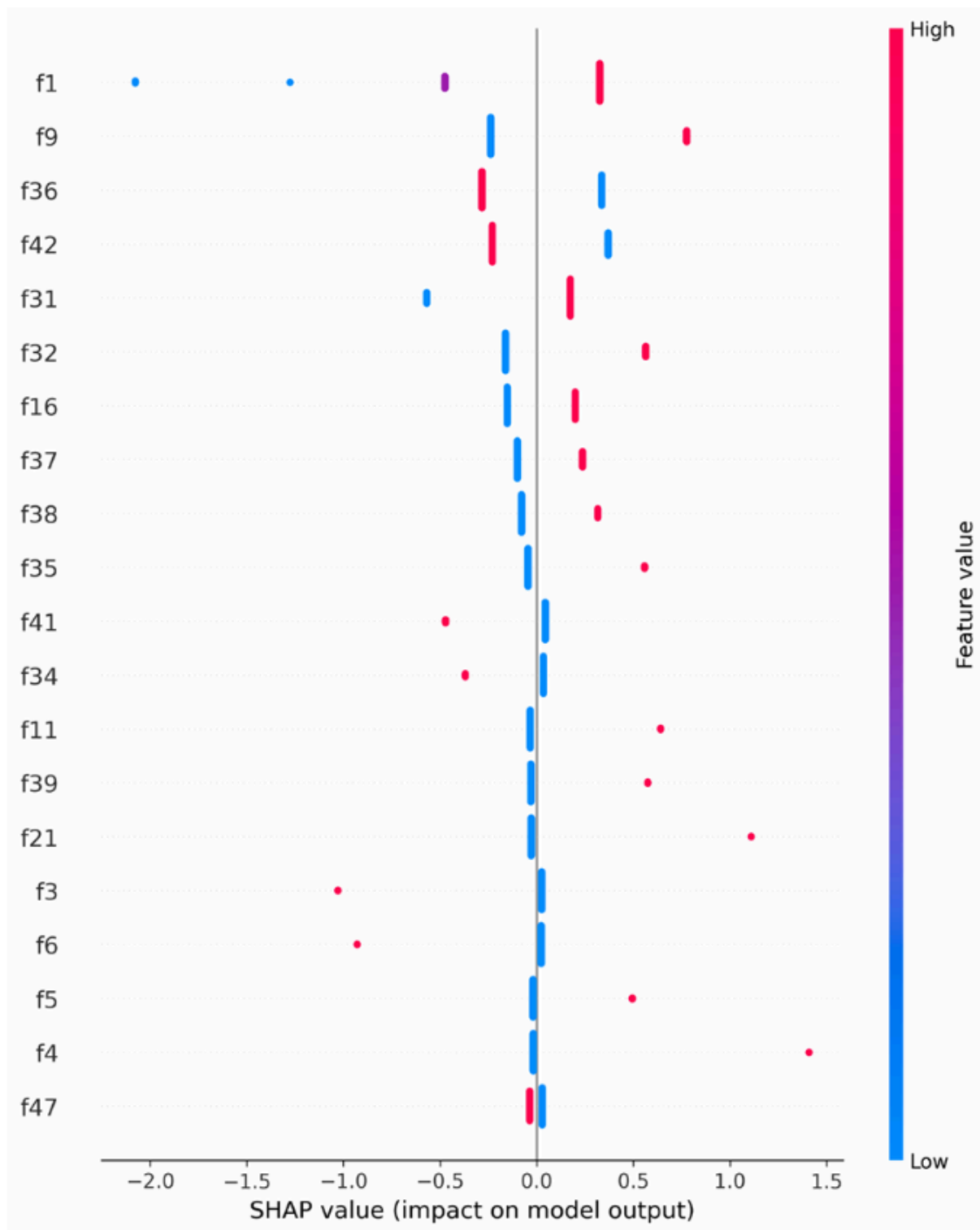


Figure 4

SHAP Summary Plot of Feature Importance for Public Health Trust Segmentation.

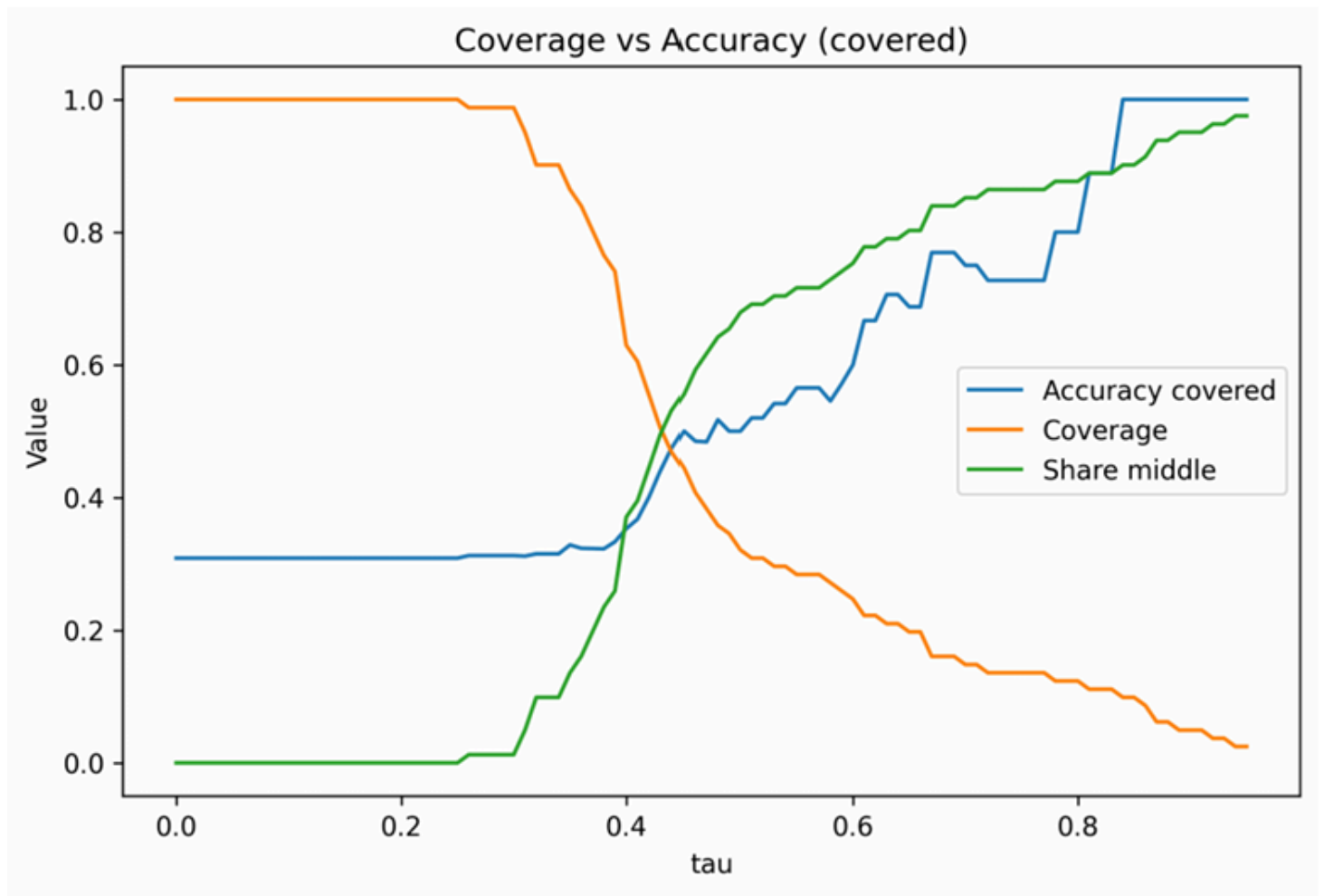


Figure 5

Impact of Confidence Threshold (τ) on Population Coverage and Segment Distribution.

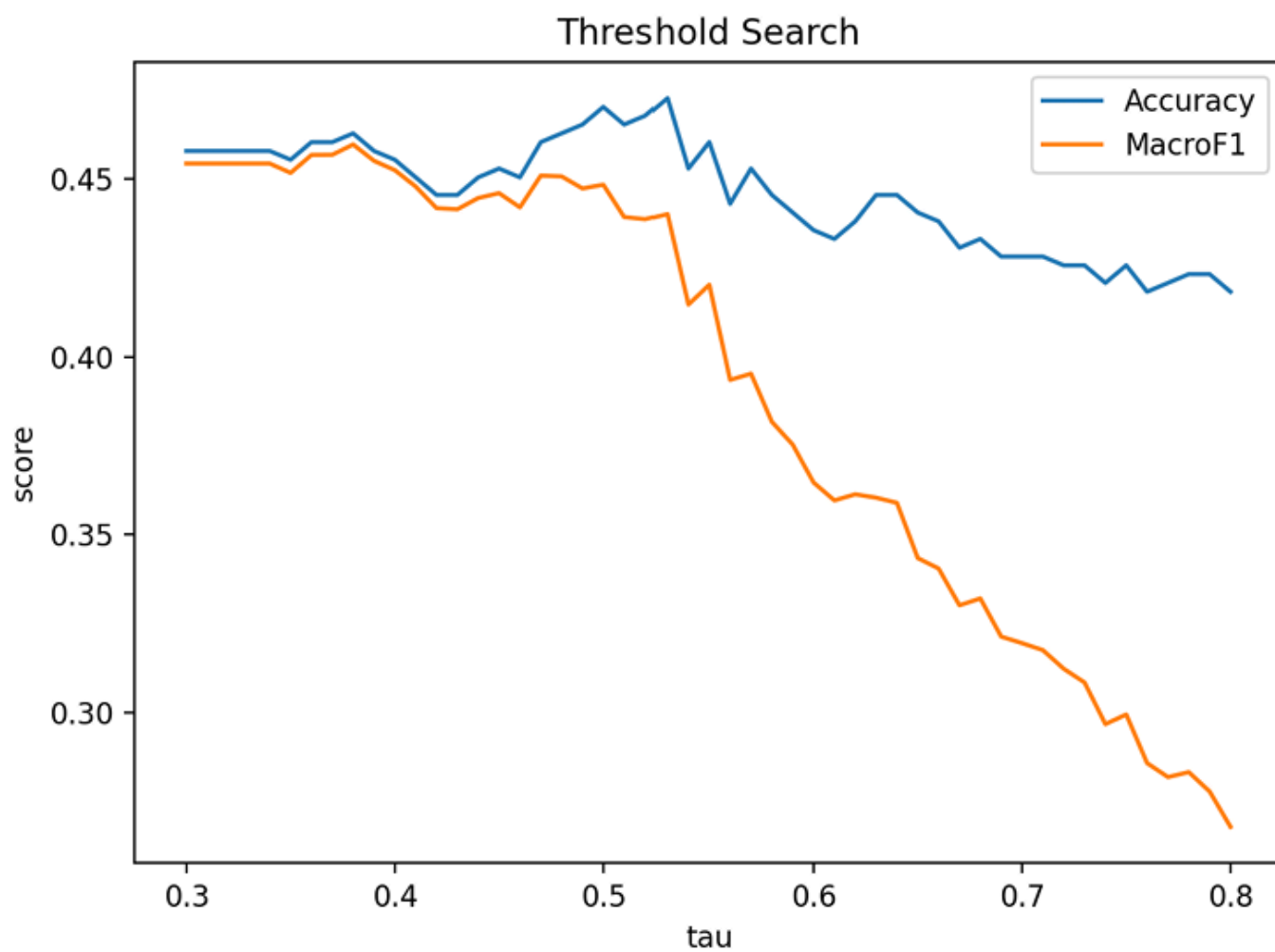


Figure 6

Model Performance Metrics as a Function of the Confidence Threshold (τ).