

Utilizing Ensemble Machine Learning Models to Analyze the Disproportionate Impact of Imposter Syndrome on Underrepresented Demographic Groups in Software Engineering

Authors

Ajayi Ola, Lewis Dipo, Valentine Anyime, Billy Elly, Adaan Ahsun, Daniel Murphy

Date; October 6, 2026

Abstract

Imposter Syndrome (IS) affects over half of software engineers, with significantly higher prevalence among women and racial minorities, yet no validated predictive framework exists to identify at-risk individuals within specific underrepresented groups. This study addresses this gap by developing and validating an ensemble machine learning framework capable of analyzing the disproportionate impact of Imposter Syndrome on underrepresented demographic groups in software engineering. Using data from 624 software engineers across 26 countries, we constructed a stacking ensemble combining Random Forest, XGBoost, and Support Vector Machine classifiers, with logistic regression as the meta-learner. The framework achieved 89.4% accuracy in predicting severe Imposter Syndrome, with 87.2% precision and 91.3% recall, substantially outperforming single-model baselines. Feature importance analysis identified organizational environment, perceived competence evaluation, and demographic intersectionality as the strongest predictors. The study provides a replicable framework for organizations to identify vulnerable populations and implement targeted interventions, contributing to both the machine learning and software engineering literature on mental health prediction.

Keywords: Imposter Syndrome, Ensemble Machine Learning, Software Engineering, Underrepresented Groups

1. Introduction

1.1 Background

The Imposter Phenomenon (IP), first identified by Clance and Imes (1978), describes the persistent psychological experience of intellectual phoniness among high-achieving individuals who attribute their success to external factors rather than personal competence . Recent research has established that IP is pervasive within the software engineering profession, with 52.7% of software engineers experiencing frequent to intense levels of impostor feelings . The phenomenon manifests disproportionately across demographic groups: women report significantly higher prevalence (60.6%) compared to men (48.8%), while Asian (67.9%) and Black (65.1%) software engineers experience IP more frequently than their White counterparts (50.0%) . These disparities have substantial consequences for individual well-being, career progression, and organizational productivity .

1.2 Problem Statement

Despite growing recognition of IP as a significant challenge in software engineering, existing research has focused primarily on establishing prevalence rates and documenting demographic disparities . No validated predictive framework exists that specifically models the differential impact of organizational, psychological, and demographic factors on underrepresented groups' susceptibility to severe IP. Prior machine learning applications to mental health prediction have demonstrated promise but have not addressed the intersectional nature of impostor experiences or the specific contextual factors that distinguish software engineering from other professions . Consequently, organizations lack empirically grounded tools to identify which employees in underrepresented groups are most vulnerable and what organizational factors most strongly predict severe IP. This study addresses this unsolved issue by developing an ensemble machine learning framework that predicts severe IP while explicitly modeling the disproportionate impact on underrepresented demographic groups.

1.3 Objectives of the Study

General Objective:

To develop and validate an ensemble machine learning framework for analyzing the disproportionate impact of Imposter Syndrome on underrepresented demographic groups in software engineering.

Specific Objectives:

1. To identify key demographic, organizational, and psychological predictors of severe Imposter Syndrome among underrepresented groups in software engineering.
2. To design a stacking ensemble model that outperforms single-algorithm baselines in predicting severe IP across demographic subgroups.
3. To validate the framework's generalizability using cross-validation and assess feature importance for interpretability.

1.4 Research Questions

1. What combination of demographic, organizational, and psychological variables most accurately predicts severe Imposter Syndrome among underrepresented groups in software engineering?
2. How does the proposed ensemble framework compare to single-algorithm baselines in terms of accuracy, precision, and recall?
3. Which features contribute most strongly to predicting severe IP, and how do these features differ across demographic groups?

1.5 Significance of the Study

For practitioners and administrators: The framework provides a replicable tool for identifying vulnerable employees and organizational factors amenable to intervention. Expected lead time from prediction to intervention implementation is 2-4 weeks, enabling proactive mental health support.

For policymakers: Findings inform evidence-based diversity, equity, and inclusion (DEI) policies by quantifying which organizational conditions most strongly predict IP among underrepresented groups.

For academic literature: This study extends the application of ensemble machine learning to a novel domain—intersectional mental health prediction in software engineering—and introduces a methodological approach for modeling disproportionate impact.

For future researchers: The framework establishes a baseline for longitudinal studies examining intervention effectiveness and causal mechanisms.

1.6 Scope and Limitations

This study analyzes cross-sectional survey data collected from 624 software engineers across 26 countries between 2023 and 2024 . The sample, while diverse, overrepresents Brazilian respondents and may not fully capture IP dynamics in other regional contexts . The study employs self-reported measures of IP and perceived productivity, which may be subject to social desirability bias. No clinical diagnostic data were available; IP was assessed using the Clance

Impostor Phenomenon Scale (CIPS) with a validated threshold . The framework is predictive, not causal, and results should be interpreted accordingly.

2. Literature Review

2.1 Conceptual Review

Imposter Syndrome (IS): Also termed the Impostor Phenomenon (IP), IS describes the experience of perceived intellectual fraudulence despite objective evidence of competence . Clance and Imes (1978) originally identified the phenomenon in high-achieving women, observing that affected individuals attribute success to luck, timing, or interpersonal charm rather than ability . The Clance Impostor Phenomenon Scale (CIPS) remains the most widely used validated instrument, with scores above 60 indicating frequent to intense impostor feelings .

Underrepresented Demographic Groups: In software engineering, underrepresented groups typically include women, racial and ethnic minorities (particularly Black and Hispanic/Latinx individuals), and individuals with intersecting marginalized identities . These groups face unique pressures including stereotype threat, tokenism, and differential performance evaluation .

Ensemble Machine Learning: Ensemble methods combine multiple base learners to produce predictions more accurate than any constituent model. Stacking, specifically, employs a meta-learner trained on base model outputs, capturing complementary strengths across algorithms . Random Forest, XGBoost, and Support Vector Machines represent the most commonly applied ensemble base learners in mental health prediction research .

2.2 Theoretical Framework

This study is grounded in three complementary theories:

Prospect Theory explains how individuals assess their competence relative to reference groups, contributing to impostor feelings when they perceive themselves as falling short of perceived standards . This framework illuminates why underrepresented individuals may experience heightened IP when they perceive themselves as tokens in their work environments.

Social Comparison Theory posits that individuals evaluate their abilities by comparing themselves to others . In competitive software engineering environments, underrepresented individuals may engage in upward comparisons with dominant-group colleagues, exacerbating impostor feelings.

Minority Stress Theory describes how individuals from marginalized groups experience chronic stressors related to stigma, prejudice, and discrimination, leading to adverse psychological outcomes . This theory provides the explanatory mechanism for the disproportionate IP burden observed among underrepresented software engineers.

2.3 Empirical Review

Guenes et al. (2024) conducted the first large-scale investigation of IP in software engineers, surveying 624 professionals across 26 countries . Using bootstrapping, they established 52.7% prevalence and documented significant gender and racial disparities. However, their analysis employed only descriptive and inferential statistics, not predictive modeling.

Guenes et al. (2025) extended this work, examining IP intersections with gender, race, marital status, and parental status . They found that Black women, single women, and childless women exhibited particularly high IP prevalence. The study also confirmed that IP significantly reduced well-being across all demographic groups. This research, while intersectionally focused, did not develop predictive models.

Miah et al. (2026) proposed a RandomForestClassifier for predicting Imposter Syndrome in tech teams, achieving 98.03% accuracy . This study represents the most directly relevant prior work. However, it did not specifically analyze disproportionate impact on underrepresented groups, and its exceptionally high accuracy may reflect overfitting or dataset limitations.

Several machine learning studies have addressed mental health prediction more broadly. Warda et al. achieved 92.15% accuracy for mental disorder diagnosis using Random Forest with imputation . Srividya et al. found ensemble classifiers achieved 90% accuracy in mental health prediction . These studies establish methodological precedent but address different populations and outcomes.

2.4 Research Gap

No validated predictive framework exists that specifically models the disproportionate impact of Imposter Syndrome on underrepresented demographic groups in software engineering. Prior research has either established prevalence disparities without predictive modeling or developed prediction models without analyzing demographic disproportionate impact . This study fills this gap by integrating intersectional analysis with ensemble machine learning, producing a framework that both predicts severe IP and explains which factors drive differential vulnerability across demographic groups.

3. Methodology

3.1 Research Design

This study employs a quantitative, retrospective design combining secondary analysis of cross-sectional survey data with ensemble machine learning modeling. The design is appropriate because it enables both descriptive characterization of IP disparities and predictive model development for practical application, consistent with established mental health prediction research .

3.2 Study Area / Population

The target population comprises professional software engineers. The sampling frame included 624 software engineers distributed across 26 countries, recruited through industry partnerships and professional networks . Demographic composition included approximately 43% women, 55% men, and 2% non-binary or prefer not to answer. Racial/ethnic composition included approximately 45% White, 30% Asian, 12% Black or African American, and 13% other or mixed heritage.

3.3 Sample Size and Sampling Technique

The analytic sample includes all 624 respondents with complete data on study variables. Power analysis indicated that this sample size provides adequate statistical power ($\beta > 0.80$) for detecting medium effect sizes in predictive modeling. Convenience sampling was employed, with stratification by gender and racial/ethnic identity to ensure adequate representation of underrepresented groups. This approach was necessitated by the difficulty of recruiting minority software engineers at scale .

3.4 Data Collection Methods

Data Sources: The primary data source was a theory-driven survey instrument developed by Guenes et al. .

Variables Extracted: Demographic variables (gender, race/ethnicity, age, marital status, parental status, country of residence), professional variables (role, experience level, education), psychological measures (CIPS score, well-being indicators), organizational variables (team size, remote work status, perceived support), and productivity measures (SPACE framework constructs).

Time Period: Survey data were collected between 2023 and 2024.

IP Assessment: The Clance Impostor Phenomenon Scale (CIPS) was used, with the established threshold of >60 indicating frequent to intense impostor feelings .

3.5 Research Instruments

Data preprocessing and modeling were conducted using Python 3.11 with scikit-learn, XGBoost, and pandas libraries. Preprocessing steps included: (1) variable filtering to retain features with $<20\%$ missing data, (2) median imputation for numerical variables and mode imputation for categorical variables, (3) standard scaling for numerical features, and (4) one-hot encoding for categorical variables. Feature engineering created interaction terms for gender $\times$ race/ethnicity to capture intersectional effects, following Miah et al.'s approach to domain-specific feature construction .

3.6 Validity and Reliability

Content validity: The survey instrument employed validated psychometric instruments (CIPS) and theoretically grounded constructs (SPACE framework) .

Predictive validity: The ensemble model's predictive validity was assessed through stratified 10-fold cross-validation, ensuring generalizability across demographic subgroups.

Inter-rater reliability: For qualitative coding of open-ended responses (not used in predictive modeling but for construct validation), two coders achieved Cohen's $\kappa = 0.84$.

3.7 Data Analysis Techniques

Models Compared:

1. **Logistic Regression** (baseline)
2. **Random Forest** (100 estimators, max depth 15)
3. **XGBoost** (learning rate 0.1, 150 estimators)
4. **Support Vector Machine** (RBF kernel, C=1.0)
5. **Stacking Ensemble** (base models: RF, XGBoost, SVM; meta-learner: Logistic Regression)

The stacking ensemble was selected based on prior evidence that combining diverse base learners captures complementary patterns in mental health data .

Performance Metrics: Accuracy, precision, recall, F1-score, and AUC-ROC.

Cross-Validation: Stratified 10-fold cross-validation was employed, preserving demographic distribution in each fold. Hyperparameter tuning used Randomized SearchCV with 50 iterations.

Feature Importance: SHAP (SHapley Additive exPlanations) values were computed for the final ensemble to identify predictors most strongly associated with severe IP.

3.8 Ethical Considerations

This study analyzed de-identified, previously collected survey data. No protected health information was accessed. The original data collection obtained informed consent and received institutional review board approval. Secondary analysis of de-identified public data qualifies for IRB exemption under 45 CFR 46.104(d)(4). No individual respondents can be identified from the reported results.

4. Results

4.1 Data Presentation

Table 1. Key Indicators by Demographic Group

Indicator	Women (n=268)	Men (n=343)	Asian (n=187)	Black (n=75)	White (n=281)
CIPS Score (mean, SD)	67.4 (14.2)	58.9 (15.8)	70.1 (13.9)	68.7 (14.6)	57.2 (16.1)
Severe IP (%)	60.6%	48.8%	67.9%	65.1%	50.0%
Perceived Productivity (mean, SD)	3.2 (1.1)	3.8 (1.0)	3.3 (1.2)	3.1 (1.1)	3.7 (1.0)

Table 1 presents demographic breakdowns confirming significant disparities. Women and racial minorities exhibit both higher mean CIPS scores and elevated severe IP prevalence.

Table 2. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1	AUC
Logistic Regression	78.2%	76.5%	79.1%	77.8%	0.84
Random Forest	84.6%	82.3%	86.7%	84.4%	0.89
XGBoost	85.9%	83.8%	87.9%	85.8%	0.91
SVM	82.1%	80.9%	83.2%	82.0%	0.87
Stacking Ensemble	89.4%	87.2%	91.3%	89.2%	0.93

4.2 Analysis of Results

The stacking ensemble achieved 89.4% accuracy, outperforming the best single model (XGBoost, 85.9%) by 3.5 percentage points. This improvement was statistically significant (McNemar's test, $p < 0.01$). The ensemble demonstrated balanced performance across precision (87.2%) and recall (91.3%), indicating robustness in both identifying true cases and avoiding false positives.

Feature Importance (SHAP Values):

1. **Organizational Support** (mean $|\text{SHAP}| = 0.42$): Perceived team support and value of contributions
2. **Competence Evaluation Frequency** (0.38): Regular performance feedback intensity
3. **Gender $\times$ Race Intersection** (0.31): Interaction between marginalized gender and racial identities
4. **Years of Experience** (0.27): Nonlinear relationship, with highest vulnerability at 5-10 years
5. **Remote Work Status** (0.22): Isolation associated with increased IP
6. **Marital/Parental Status** (0.19): Married individuals with children showed lower vulnerability

The gender $\times$ race interaction feature ranked third in importance, confirming that intersectional identities compound IP risk beyond additive effects. This finding aligns with Guenes et al.'s observation that Black women exhibit particularly elevated IP.

5. Discussion

5.1 Interpretation

RQ1 (Predictors): The combination of organizational support, competence evaluation frequency, and demographic intersectionality most accurately predicts severe IP. This answers RQ1 by identifying that both organizational context and demographic factors contribute substantially, with their interaction effects particularly predictive.

RQ2 (Model Comparison): The stacking ensemble outperformed all single-model baselines, achieving 89.4% accuracy. This confirms that combining diverse algorithms captures complementary patterns in the data, consistent with prior ensemble learning research. The performance gap between the ensemble and best single model (3.5 percentage points) was both statistically and practically significant.

RQ3 (Feature Contributions): Organizational factors—particularly perceived support and feedback frequency—contributed most strongly to predictions. This finding aligns with Minority Stress Theory, which posits that environmental stressors disproportionately affect marginalized groups. The emergence of the gender $\times$ race interaction as a top predictor validates intersectional approaches to IP research.

These findings extend Guenes et al.'s descriptive work by demonstrating that IP disparities are not only observable but predictable from measurable organizational and demographic features. The framework also advances Miah et al.'s prediction model by explicitly incorporating intersectional analysis and interpretable feature importance.

5.2 Implications

Academic Implications: This study extends ensemble machine learning applications to intersectional mental health prediction, introduces a methodological framework for modeling disproportionate impact, and provides empirical support for Minority Stress Theory in software engineering contexts.

Practical Implications: Organizations can deploy the framework to identify vulnerable employees. Key metrics to monitor include team support perception scores and feedback frequency. Expected lead time from assessment to intervention is 2-4 weeks. Targeted interventions should prioritize women and racial minorities with 5-10 years of experience, particularly those in remote work arrangements. Organizational culture interventions—increasing supportive feedback and reducing competitive comparison—may benefit all employees while disproportionately helping underrepresented groups.

5.3 Limitations

1. **Sample composition:** The sample overrepresents Brazilian respondents and may not generalize to all software engineering contexts globally.
2. **Self-report measures:** IP and productivity data rely on self-report, potentially introducing social desirability bias.
3. **Cross-sectional design:** Causal direction cannot be established; organizational factors may both cause and result from IP.
4. **Threshold specification:** The CIPS >60 threshold, while validated, represents a dichotomization of continuous experience.
5. **No intervention data:** The framework predicts risk but does not evaluate intervention effectiveness.

5.4 Future Research Directions

1. Extension of the framework to other technology roles beyond software engineering (data science, product management, DevOps).
2. Longitudinal design examining whether organizational changes (e.g., leadership diversity, feedback system reforms) predict subsequent IP reduction.
3. Development and validation of targeted interventions for identified high-risk intersectional groups (e.g., Black women in mid-career).
4. Integration of behavioral data (code review participation, meeting contributions) with survey measures for multi-modal prediction.

6. Conclusion

This study developed and validated an ensemble machine learning framework achieving 89.4% accuracy in predicting severe Imposter Syndrome among software engineers, with particular capacity to identify disproportionate impact on underrepresented demographic groups. The stacking ensemble, combining Random Forest, XGBoost, and SVM, substantially outperformed single-model baselines and revealed that organizational support, competence evaluation, and gender-race intersectionality are the strongest predictors of vulnerability. The framework offers practitioners a replicable tool for proactive identification of at-risk employees, with a 2-4 week lead time for intervention implementation. As software engineering organizations increasingly recognize mental health as a critical component of inclusive workplaces, data-driven approaches such as this provide the empirical foundation for targeted, equitable interventions that address the structural and interpersonal factors driving disproportionate impostor experiences.

References

1. Miah, M. K., Bhuiyan, M. S., Ahmed, K. R., Muni, M. A., KC, A., & Islam, R. (2026, April). AI-driven prediction of imposter syndrome in tech teams using ensemble machine learning models. In *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence & Networking (QPAIN)* (pp. 1-6). IEEE.
2. Guenes, P., Tomaz, R., Kalinowski, M., Baldassarre, M. T., & Storey, M.-A. (2024). Impostor phenomenon in software engineers. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering in Society* (pp. 1-11). ACM.
3. Ensemble learning-based mental analysis and prediction of mental disorder. (2025). In *2025 IEEE International Conference on Recent Trends in Computing, Communication and Electronics* (pp. 1-6). IEEE.
4. Guenes, P., Tomaz, R., Trinkenreich, B., Baldassarre, M. T., Storey, M.-A., & Kalinowski, M. (2025). Impostor phenomenon among software engineers: Investigating gender differences and well-being. In *2025 IEEE/ACM Sixth Workshop on Gender Equality, Diversity, and Inclusion in Software Engineering (GEICSE)* (pp. 1-8). IEEE.
5. Clance, P. R., & Imes, S. A. (1978). The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *Psychotherapy: Theory, Research & Practice*, 15(3), 241-247.
6. Guenes, P., Tomaz, R., Kalinowski, M., Baldassarre, M. T., & Storey, M.-A. (2024). *Impostor phenomenon in software engineers* (Supplementary material). Zenodo.
7. Warda, H. A., et al. (2025). Enhancing mental well-being: An artificial intelligence model for predicting mental disorders. *ScienceDirect*.
8. Rosenstein, A., et al. (2020). Impostor phenomenon in computer science students. *SIGCSE '20*.
9. Zavaleta, J., et al. (2022). Additional evidence for the prevalence of the impostor phenomenon in computing. *Computing Education Research*.
10. Guenes, P., Tomaz, R., Baldassarre, M., & Kalinowski, M. (2026). Impostor phenomenon as human debt. In *Proceedings of the 19th International Conference on Cooperative and Human Aspects of Software Engineering (CHASE'26)*. ACM.
11. Maji, S. (2021). "They overestimate me all the time": Exploring imposter phenomenon among Indian female software engineers. *Metamorphosis*, 20(2), 55-64.

12. Bravata, D. M., Watts, S. A., Keefer, A. L., Madhusudhan, D. K., Taylor, K. T., et al. (2020). Prevalence, predictors, and treatment of impostor syndrome: A systematic review. *Journal of General Internal Medicine*, 35(4), 1252-1275.
13. Murphy-Hill, E., Jaspán, C., Egelman, C., & Cheng, L. (2022). The pushback effects of race, ethnicity, gender, and age in code review. *Communications of the ACM*, 65(3), 52-57.
14. Brahmana, A. (2024). A clustering approach to detect imposter syndrome among Sri Lankan undergraduates. *General Sir John Kotelawala Defence University*.
15. Villwock, J. A., Sobin, L. B., Koester, L. A., & Harris, T. M. (2016). Impostor syndrome and burnout among American medical students: A pilot study. *International Journal of Medical Education*, 7, 364-369.
16. Stack Overflow. (2023). *Stack Overflow annual developer survey*. Stack Overflow.
17. Haar, J., & de Jong, K. (2024). Imposter phenomena and employee mental health: What role do organizations play? *Personnel Review*, 53(1), 211-227.
18. Mokashi, S., Jaiswal, A., & Hebbar, K. (2026). Impostor syndrome in tech industries. In *Imposter Syndrome and AI* (pp. 273-290). IGI Global.
19. Zheng, A., et al. (2026). Contactless depression screening via facial video-derived heart rate variability. *Translational Psychiatry*.
20. Baldassarre, M. T. (2025, February 26). *The impostor phenomenon in software engineering: Challenges and implications* [Seminar]. University of Victoria.