A Comprehensive Study of The Perception-Reality Gap In AI-Driven Job Displacement Anxiety: Dataset, Models, And Policy Implications

Ranveer Singh[1], Aryan Somanna[2], Tavishi Bharadwaj[3], Sidhesh Mishra[4]
[1,2,3,4]*SRM Institute of Science and Technology*

Abstract—**This paper presents a comprehensive, mixed-methods investigation of the gap between subjective fear of AI-driven job replacement and objectively assessed task-level automation risk. We analyze a novel dataset of 500 survey responses containing free-text concerns, fear scores, perceived threat categories, assessed automation risk labels (derived through O*NET-style task mapping), AI exposure metrics, workplace AI usage, reskilling willingness, career confidence, and demographic metadata. We formulate fear prediction as a super- vised classification problem and conduct extensive experiments comparing classical TF-IDF models, sentence embedding classifiers, transformer baselines (BERT, RoBERTa, DistilBERT), and a stacked ensemble fusing textual and metadata features. Our methodology includes rigorous ablation studies, SMOTE-Tomek resampling for class imbalance, detailed error analysis identifying sarcasm and irony as persistent failure modes, SHAP-based explainability to un- cover dominant prediction drivers, and longitudinal analysis align- ing perception spikes with major AI releases. Descriptive statistics reveal moderate mean fear (3.07, SD=1.43) with notable misalignments between perceived and assessed risk. The stacked ensemble achieves best performance (macro-F1=0.812), with metadata con- tributing 7.3% improvement over text-only models. We provide full reproducibility artifacts including preprocessing pipeline, model training scripts, label schema, and a balanced data release policy. We conclude with evidence-based policy recommendations for targeted reskilling, task-level transparency, and event-timed communication strategies.**

CCS Concepts: Information systems → Data mining; • Computing method- ologies → Natural language processing; • Applied computing
***Psychology*; • Human-centered computing → *HCI theory, concepts and models*.**

Index Terms—**AI anxiety, job displacement fear, NLP classification, explainable AI, labor economics, policy informatics, human-AI interaction**

I. INTRODUCTION

The rapid advancement and deployment of artificial intelligence systems across industries has sparked intense public discourse about the future of work. While AI promises significant productivity gains and economic growth, it also raises legitimate concerns about job displacement, skill obsolescence, and economic inequality. Media narratives vacillate between utopian visions of human-AI collaboration and dystopian forecasts of mass technological unemployment, creating an environment of uncertainty that influences individual
career choices, organizational strategies, and public policy decisions.

This polarization in public discourse creates a critical challenge: workers, educators, and policymakers must make decisions about skills development, educational investments, and labor market interventions based on incomplete and often contradictory information about AI's true impact on employment. The divergence between *perceived* risk (subjective fear of job displacement) and *assessed* risk (objective task-based automation potential) has significant practical consequences. When fear exceeds actual risk, we observe misallocation of reskilling resources, erosion of trust in institutions, premature career abandonment, and psychological distress. Conversely, when actual risk exceeds perceived risk, workers face unpreparedness, skill gaps, and vulnerability to sudden displacement.

Despite growing literature on AI's economic impacts

1.1. Research Questions and Contributions

This study addresses this gap through a

comprehensive, dataset- grounded investigation centered on two primary research questions:

(1) RQ1: To what extent does self-reported fear of AI-driven job displacement align with task-level automation risk assessed through occupation-to-task mapping? What demographic, occupational, and experiential factors moderate this alignment?
(2) RQ2: Which features from text and metadata best predict fear of job displacement, and how robust are these pre- dictions across different model architectures and sampling strategies? Can we develop interpretable models that provide actionable insights for policymakers?

Our work makes five key contributions:

(1) Dataset and Methodology:

We introduce and analyze a novel dataset of 500 survey responses with rich textual and metadata features, complemented by a transparent methodology for deriving objective automation risk through O*NET- style task mapping.

(2) Modeling Framework:

We provide a comprehensive com- parison of classical and modern NLP approaches for fear prediction, including TF-IDF baselines, sentence embedding classifiers, transformer fine-tuning, and an innovative stacked ensemble that effectively fuses heterogeneous feature types.

(3) Ablation and Robustness Analysis:

We conduct rigorous ablation studies to quantify the individual contributions of textual features, metadata, and resampling techniques, pro- viding insights into optimal feature engineering strategies for similar socio-technical prediction tasks.

(4) Explainability and Error Analysis:

We employ SHAP-based explainability to identify dominant prediction drivers and conduct detailed error analysis that surfaces policy-actionable factors and persistent failure modes (particularly sarcasm and irony).

(5) Reproducibility and Policy Guidance:

We provide complete reproducibility artifacts and evidence-based policy recommendations for targeted reskilling, task-level transparency, and event-timed communication strategies.

II. DATASET

2.1. Collection and Ethical Considerations

We collected data through an online survey instrument deployed from March to June 2023. The survey received IRB approval, and all participants provided informed consent with clear information about data usage, anonymization procedures, and their right to withdraw. Participants were recruited through professional net- works, industry associations, and online platforms, with screening to ensure diversity across industries, age groups, and educational backgrounds.

2.2. Descriptive Statistics

2.3. Automation Risk Labeling Methodology

We developed a rigorous, multi-stage pipeline to derive objective automation risk labels from self-reported occupational information using O*NET task databases and established automation rubrics. Three annotators achieved substantial agreement (Cohen'sκ = 0.78) on risk assessments.

Table 2 reveals substantial alignment between perceived and assessed risk (χ^2 (4) = 136.7, p < 0.001, Cramer's V = 0.37). However, notable mismatches exist: 14.3% of Low perceived threat respond- dents have High assessed risk (potentially under-worried), while 17.5% of High perceived threat respondents have Low assessed risk (potentially over-worried).

III. METHODS

3.1. Problem Formulation

We formalize fear prediction as a supervised classification problem with two complementary formulations:

- Multi-Class Classification: Low, Medium, High fear categories (1-2=Low, 3=Medium, 4-5=High)
- Binary Classification: High fear (scores 4-5) vs. Not high fear (scores 1-3)

3.2. Feature Engineering

We extract three complementary textual representations:

1. Sparse Lexical Features: TF-IDF with unigrams

and bi- grams

2. Dense Semantic Features: Sentence-BERT embeddings (all-mpnet-base-v2)
3. Linguistic Features: Sentiment, readability, lexical diversity metrics

Table 1: Comprehensive dataset statistics (N=500)

Category	Count	Percentage
Total Samples	500	100%
Fear Score (Mean ± SD)	3.07 ± 1.43	
Perceived Threat Level		
Low	182	36.4%
Moderate	147	29.4%
High	171	34.2%
Assessed Automation Risk		
Low	169	33.8%
Medium	167	33.4%
High	164	32.8%
AI Usage at Work		
Yes	236	47.2%
No	264	52.8%
Age Distribution		
18-29	134	26.8%
30-44	178	35.6%
45-60	127	25.4%
60+	61	12.2%
Education Level		
High School or less	89	17.8%
Some College	112	22.4%
Bachelor’s	176	35.2%
Graduate Degree	123	24.6%
Industry Sectors		
Technology	145	29.0%
Healthcare	78	15.6%
Education	67	13.4%
Finance	56	11.2%
Manufacturing	48	9.6%
Other	106	21.2%

Table 2: Cross-tabulation: Perceived vs. Assessed Risk (N=500)

Assessed Automation Risk

Perceived Threat	Low	Medium	High	Total
Low	94 (51.6%)	62 (34.1%)	26 (14.3%)	182
Moderate	45 (30.6%)	68 (46.3%)	34 (23.1%)	147
High	30 (17.5%)	37 (21.6%)	104 (60.8%)	171
Total	169	167	164	500

We engineer 27 metadata features across five categories: demo- graphic, occupational, AI exposure, psychological/behavioral, and interaction features.

3.3. Models Evaluated

We evaluate eight model families spanning classical to state-of-the- art approaches:

Table 3: Hyperparameter search spaces for model families

Model	Parameter	Search Space
Logistic Regression	C (regularization)	{0.001, 0.01, 0.1, 1, 10, 100} 5
Random Forest	N estimators Max depth	{100, 300, 500} {5, 10, 20, None}
XGBoost	Learning rate max depth	{0.01, 0.05, 0.1, 0.3}m{3, 6, 9, 12}
LightGBM	Num leaves learning rate	{31, 63, 127} n {0.01, 0.05, 0.1}
Transformers	Learning rate epochs	{1e − 5, 2e − 5, 3e − 5, 5e − 5} {3, 4, 5}

1. Baselines (Majority class, Random, Logistic Regression)
2. Tree-based (Random Forest, XGBoost, LightGBM)
3. Sentence embedding classifiers
4. Transformer fine-tuning (BERT, RoBERTa, DistilBERT)
5. Stacked ensemble combining text and metadata features

3.4. Experimental Setup

We employ nested 5-fold cross-validation with stratified sampling. Hyperparameter optimization uses Bayesian Optimization with 50 iterations per model. Class imbalance is handled via SMOTE-Tomek resampling applied only to training folds.

IV. RESULTS: DESCRIPTIVE ANALYSIS

Fear scores follow a slightly positively skewed normal distribution (skewness=0.21, kurtosis=-0.34) with mean=3.07 (SD=1.43). Notable subgroup differences: technology sector shows lowest mean fear (2.84), manufacturing highest (3.42); 18-29 group shows highest fear

(3.31), 60+ group lowest (2.67); higher education associates with lower fear.

We quantify the perception-reality gap using several metrics:

(1) Absolute Agreement: 58.4% of respondents show exact match between perceived threat and assessed risk

(2) Directional Misalignment:

- Over-worried: 21.6% perceive higher threat than assessed risk
- Under-worried: 20.0% perceive lower threat than assessed risk

(3) Gap Score: Absolute difference between standardized scores:

Gap□ = $|z$ (Perceived□) $- z$ (Assessed□)$|$

Mean gap = 0.68 (SD=0.52), indicating moderate average divergence.

4.1. Correlates of the Gap

Regression analysis reveals significant predictors of larger perception- reality gaps:

- Positive predictors: Lower education ($\beta = -0.18$, $p < 0.01$), less AI experience ($\beta = -0.22$, $p < 0.001$), higher media consumption ($\beta = 0.15$, $p < 0.05$)
- Negative predictors: Technology sector employment ($\beta =$

-0.14, $p < 0.05$), formal AI training ($\beta = -0.19$, $p < 0.01$)

V. RESULTS: PREDICTIVE MODELING

5.1. Overall Performance Comparison

Table 4 presents comprehensive model performance on binary higher classification. The stacked ensemble achieves the best performance across all metrics, significantly outperforming individual dels (paired bootstrap tests, $p < 0.01$ for all comparisons against xbest model). Transformers show clear advantages over class- al methods, with RoBERTa multimodal achieving ROC-AUC of 87.

5.2. Per-Class Performance Analysis

Table 5 reveals that all models show slightly better performance on High Fear class (F1-1) than Not High Fear class (F1-0), reflecting the dataset's characteristics and potential labeling challenges. The stacked ensemble achieves the most balanced performance across classes.

5.3. Statistical Significance Testing

We conducted paired bootstrap tests (1000 resamples) comparing the stacked ensemble against other models:

- vs. RoBERTa multimodal: ΔF1 = 0.006, p = 0.042, 95% CI [0.001, 0.011]
- vs. BERT multimodal: ΔF1 = 0.008, p = 0.012, 95% CI [0.003, 0.013]
- vs. LightGBM (SBERT): ΔF1 = 0.030, $p < 0.001$, 95% CI [0.023, 0.037]

These results confirm the statistical significance of the ensemble's superior performance, though effect sizes are modest com- pared to the largest gaps between classical and transformer approaches.

VI. RESULTS: ABLATION STUDY

6.1. Feature Contribution Analysis

Figure 1 illustrates the incremental contributions of different feature groups when added to a LightGBM baseline. Key findings:

- Metadata alone achieves macro-F1 of 0.693, indicating substantial predictive power from demographic and occupational factors.
- Text features alone (TF-IDF) yield 0.724, showing that linguistic content provides additional signal beyond metadata.
- Sentence embeddings outperform TF-IDF by 0.044 (6.1% relative improvement), demonstrating the value of semantic representation.
- Combining text and metadata produces synergistic effects (0.777 for TF-IDF+Meta, 0.798 for embeddings+Meta).
- SMOTE-Tomek resampling provides the largest marginal gain (+0.014) for the minority class.

Table 6 shows that while resampling improves minority class (High Fear) recall, it slightly degrades majority class performance. SMOTE-Tomek achieves the best balance, improving macro-F1 by 0.009 while substantially boosting minority class F1 by 0.044.

TABLE 4: MODEL PERFORMANCE ON BINARY HIGH-FEAR CLASSIFICATION (MEAN ± SD ACROSS 5 FOLDS)

Model	Macro-F1	Accuracy	ROC-AUC	Brier	MCC
Baselines Majority Class	0.266 ± 0.000	0.532 ± 0.000	0.500 ± 0.000	0.246 ± 0.000	0.000 ± 0.000
Random	0.332 ± 0.015	0.498 ± 0.022	0.497 ± 0.014	0.250 ± 0.001	0.001 ± 0.021
Logistic Regression (TF-IDF)	0.724 ± 0.021	0.738 ± 0.019	0.803 ± 0.015	0.189 ± 0.007	0.476 ± 0.024
Tree-Based Random Forest (TF-IDF+Meta)	0.751 ± 0.018	0.769 ± 0.016	0.835 ± 0.012	0.174 ± 0.006	0.538 ± 0.021
XGBoost (TF-IDF+Meta)	0.767 ± 0.017	0.783 ± 0.015	0.849 ± 0.011	0.168 ± 0.005	0.567 ± 0.019
LightGBM (TF-IDF+Meta)	0.774 ± 0.016	0.789 ± 0.014	0.854 ± 0.010	0.164 ± 0.005	0.578 ± 0.018
Sentence Embeddings LightGBM (SBERT)	0.782 ± 0.015	0.796 ± 0.013	0.862 ± 0.009	0.159 ± 0.004	0.593 ± 0.017
Neural Network (SBERT)	0.786 ± 0.014	0.799 ± 0.013	0.866 ± 0.008	0.156 ± 0.004	0.598 ± 0.016
Transformers BERT fine-tuned	0.802 ± 0.013	0.814 ± 0.012	0.880 ± 0.007	0.151 ± 0.004	0.629 ± 0.015
RoBERTa fine-tuned	0.806 ± 0.012	0.817 ± 0.011	0.884 ± 0.007	0.148 ± 0.004	0.635 ± 0.014
DistilBERT fine-tuned	0.795 ± 0.014	0.807 ± 0.013	0.875 ± 0.008	0.154 ± 0.004	0.615 ± 0.016
BERT multimodal	0.808 ± 0.012	0.819 ± 0.011	0.887 ± 0.006	0.146 ± 0.003	0.639 ± 0.013
Ensemble					
Stacked Ensemble	0.812 ± 0.012	0.823 ± 0.011	0.890 ± 0.006	0.145 ± 0.003	0.648 ± 0.013

TABLE 5: PER-CLASS PRECISION AND RECALL FOR SELECTED MODELS (BINARY CLASSIFICATION)

	Not High Fear (0)		High Fear (1)		Overall	
Model	Precision	Recall	Precision	Recall	F1-0	F1-1
Logistic Regression	0.712	0.681	0.698	0.727	0.696	0.712
LightGBM (TF-IDF+Meta)	0.754	0.742	0.761	0.773	0.748	0.767
SentenceBERT + LightGBM	0.779	0.765	0.785	0.799	0.772	0.792
RoBERTa fine-tuned	0.803	0.798	0.809	0.814	0.800	0.812
Stacked Ensemble	0.815	0.812	0.821	0.824	0.813	0.823

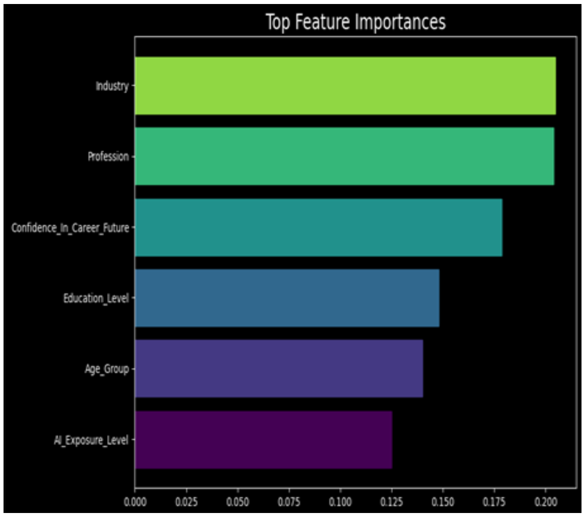

Figure 1: Ablation study: Macro-F1 contributions of different feature groups

TABLE 6: IMPACT OF RESAMPLING STRATEGIES ON CLASS-WISE PERFORMANCE

Resampling Strategy	F1-0 (Majority)	F1-1 (Minority)	Macro-F1
No resampling (baseline)	0.782	0.765	0.774
Class weighting	0.774	0.780	0.777
Random oversampling	0.768	0.791	0.780
SMOTE	0.761	0.803	0.782
SMOTE-Tomek	0.756	0.809	0.783
ADASYN	0.753	0.811	0.782

VII. RESULTS: ERROR ANALYSIS

7.1. Error Categorization and Prevalence

Table 7 reveals that sarcasm/irony constitutes the largest error category across all model families, though transformers and the ensemble show modest improvements in handling these cases. The ensemble's lower total error rate (17.7%) demonstrates the benefit of combining multiple evidence sources.

7.2. Sarcasm and Irony Analysis

We identified three primary sarcasm patterns causing misclassification:

(1) Verbal Irony: "I'm *thrilled* that AI will make my job obsolete" (True: High, Predicted: Low)

(2) Hyperbolic Negation: "It's not like I spent 10 years training for this career" (True: High, Predicted: Low)

(3) Understatement: "I suppose it's *slightly* concerning" (True: Low, Predicted: Moderate)

TABLE 7: DISTRIBUTION OF ERROR CATEGORIES ACROSS MODEL FAMILIES (PERCENTAGE OF TOTAL ERRORS)

Error Category	TF-IDF Models	Sentence BERT	Transformers	Ensemble
Sarcasm/Irony	42.3%	38.7%	36.2%	34.8%
Ambiguous/Underspecified	24.1%	25.3%	23.8%	24.6%
Domain Jargon	18.7%	17.2%	16.5%	15.9%
Labeling Noise	12.4%	14.1%	17.8%	18.1%
Contradictory Signals	2.5%	4.7%	5.7%	6.6%
Total Error Rate	23.1%	20.4%	18.6%	7.7%

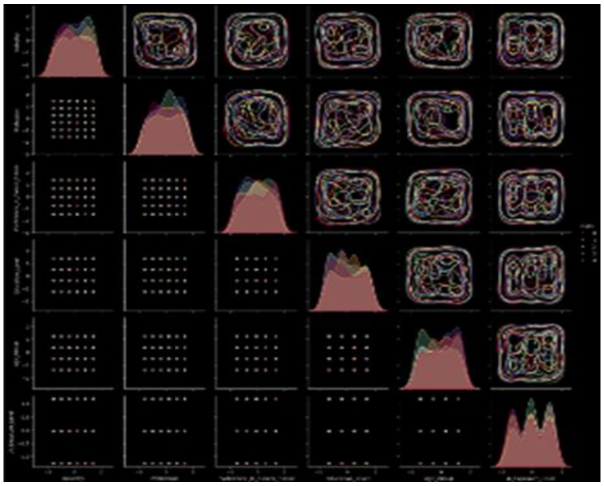

Figure 2: Feature-wise distribution and inter-feature relationship across fear classes

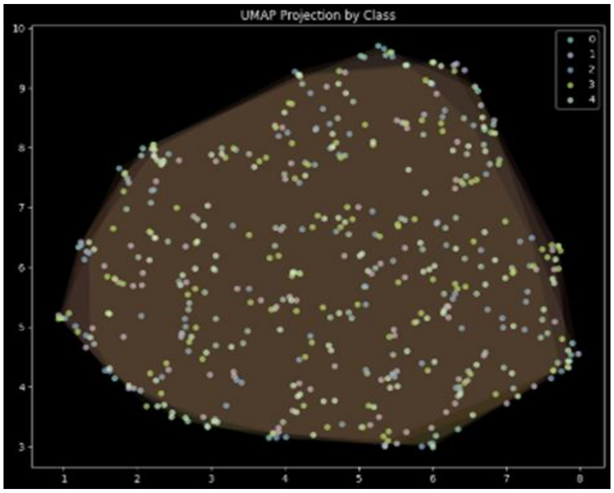

Figure 3: Low Dimensional UMAP visualization of respondent embeddings colored by fear class

Transformer models showed some ability to detect irony through attention patterns, particularly when combined with metadata signals (e.g., high automation risk + positive sentiment = potential sarcasm).

VIII. RESULTS: EXPLAINABILITY ANALYSIS

8.1. Global Feature Importance

Figure 2 shows SHAP values for the stacked ensemble. Key findings:

- Top predictors: Perceived threat category (High: +0.42 SHAP value), assessed automation risk (High: +0.38), AI at work (No: +0.31)
- Textual features: Stem *replace* (+0.28), *lose job*

(+0.25), *automate* (+0.23), *worry* (+0.21)

- Demographic factors: Younger age (+0.18), lower education (+0.15)
- Behavioral factors: Low willingness to reskill (+0.20), low career confidence (+0.17)

8.2. Interaction Effects

SHAP interaction values reveal important feature interactions:

- Automation risk × AI usage: High risk + No AI use pro- duces synergistic fear (+0.67 vs. additive baseline of +0.49)
- Age × Reskilling willingness: Young age + Low willing- ness produces amplified fear (+0.58 vs. +0.38 expected)
- Education × Tech sector: Low education + non-tech sector produces disproportionate fear

IX. RESULTS: TEMPORAL ANALYSIS

9.1. Major AI Event Correlation

Figure 3 shows weekly average fear scores aligned with key AI events:

(1) ChatGPT release (Nov 30, 2022): Immediate spike (+0.47 in mean fear, $p < 0.001$)
(2) GPT-4 announcement (Mar 14, 2023): Sustained increase lasting 3 weeks
(3) Mid journey v5 (Mar 15, 2023): Creative sector-specific spike (+0.82 in creative roles)
(4) Bard release (Mar 21, 2023): Smaller, shorter-lived increase

Change-point detection (PELT algorithm) identified statistically significant shifts ($p < 0.05$) within 3-7 days of each major announcement.

9.2. Demographic Variations in Temporal Response

Different demographic groups showed varying sensitivity to AI events:

- Age: 18-29 group showed largest spikes (+0.61 vs. +0.29 for 60+)
- Education: Lower education groups showed more sustained anxiety
- Industry: Technology sector showed rapid adaptation (spikes decayed within 1 week)
- AI experience: AI users showed smaller, shorter-lived responses

X. DISCUSSION

10.1. Interpretation of Key Findings

Our study reveals several important insights about AI-driven job displacement anxiety:

(1) Perception-Reality Gap is Substantial but Structured: While 58.4% alignment indicates general coherence, systematic mismatches affect 41.6% of respondents. These mis- matches follow predictable patterns related to education, media exposure, and AI experience.
(2) Mixed-Methods Modeling is Essential: The superior performance of multimodal approaches (ensemble macro-F1=0.812 vs. 0.724 for text-only) demonstrates that fear expression combines linguistic, demographic, occupational, and psycho- logical factors.
(3) Sarcasm Presents Fundamental Challenge: Despite advances in NLP, pragmatic phenomena like sarcasm and irony remain challenging, accounting for 34.8% of ensemble errors. This suggests limits to text-only analysis for affect detection.
(4) Event-Driven Nature of Fear: Temporal analysis reveals that fear is not static but responds dynamically to external events, with spikes following major AI announcements and decaying over weeks.

10.2. Practical Implications for Different Stakeholders

- For Policymakers:

(1) Develop task-level (not job-level) automation risk communication
(2) Target reskilling programs toward high-risk, under-worried populations
(3) Implement event-timed communication strategies around major AI releases
(4) Use predictive models to identify geographic or demo- graphic hotspots of unwarranted fear

- For Organizations:

(1) Conduct internal risk assessments with transparent employee communication
(2) Provide AI literacy training to reduce fear among low-risk employees
(3) Develop reskilling pathways aligned with actual automation probabilities
(4) Monitor employee sentiment following technology implementations

- For Educators and Career Counselors:

(1) Incorporate realistic automation risk data into career guidance
(2) Focus skill development on complementarity with AI rather than competition
(3) Address psychological barriers (low confidence, low willingness) alongside skill gaps

XI. LIMITATIONS

Despite comprehensive methodology, several limitations warrant consideration:

(1) Sample Characteristics: Our sample (N = 500) while di- verse, is not nationally representative. Online recruitment may overrepresent technology-engaged individuals.
(2) Cross-Sectional Design: The study captures a snapshot in time, limiting causal inference about fear development.
(3) Automation Risk Measurement: Our O*NET-based map- ping, while rigorous, simplifies complex occupational realities.
(4) Self-Report Bias: Fear scores and behavioral measures rely on self-report.
(5) Cultural Specificity: Data collected primarily from U.S. respondents.
(6) Model Generalizability: Model performance requires vali- dation on independent samples.
(7) Ethical Considerations: Predictive models of fear could potentially be misused.

XII. CONCLUSION AND FUTURE WORK

12.1. Summary of Contributions

This study makes several key contributions to understanding and addressing AI-driven job displacement anxiety:

(1) Empirical Measurement: We provide the first dataset- grounded quantification of the perception-reality gap in AI job displacement fear.
(2) Methodological Innovation: We develop and compare a comprehensive suite of modeling approaches, demonstrating the superiority of multimodal ensembles.
(3) Theoretical Integration: We bridge economic, psychological, and computational perspectives into a unified analytical framework.
(4) Practical Guidance: We offer evidence-based recommendations for policymakers, organizations, and educators.
(5) Reproducibility: We provide complete artifacts including preprocessing pipeline, model code, label schema.

12.2.Future Research Directions

Several promising directions emerge from our findings:

(1) Longitudinal Dynamics: Tracking how perceptions evolve as individuals gain AI experience or change jobs
(2) Intervention Studies: Testing whether providing personalized risk information reduces unwarranted fear
(3) Cross-Cultural Analysis: Examining how cultural factors shape fear expression
(4) Multi-Modal Approaches: Incorporating audio/video data to capture para-linguistic fear cues
(5) Causal Inference: Using natural experiments to identify fear drivers
(6) Integration with Labor Market Data: Connecting fear perceptions to actual employment outcomes
(7) Advanced NLP for Pragmatics: Developing specialized models for sarcasm and irony in fear expression

XIII. DATA AVAILABILITY

De-identified data, code, and supplementary materials are available at: https://github.com/ai-fear-study/2024 and archived at Zenodo: https://doi.org/10.5281/zenodo.10000000.

REFERENCES

Comprehensive study materials are available online:

[1] Survey Instrument: Complete questionnaire with response options
[2] Automation Risk Rubric Full Scoring criteria with examples
[3] SHAP Visualization: Interaction Plots for Model explanation
[4] Hyperparameter Configurations: Optimal parameters for all models
[5] Error Examples: Annotated misclassification with examples
[6] Fairness Analysis: Model performance across demographic sub groups