Managing Distribution Shift in Speech Emotion Recognition: An Empirical Study with Confidence-based Filtering

Ruwini Madhushika Herath
Independent Researcher
Stockholm, Sweden

ABSTRACT

Speech emotion recognition (SER) plays an important role in human–computer interaction, healthcare, and customer service. Yet SER models often degrade when applied across genders or to external corpora, limiting their reliability in real-world deployments. This study investigates the robustness of classical classifiers- Logistic Regression, Random Forests, and XGBoost, under gender and domain shifts, with a focus on confidence-based routing as a mitigation strategy. In-domain experiments demonstrated strong performance for tree-based ensembles, with Random Forests achieving up to 0.879 accuracy and XGBoost 0.914 on gender-specific training, while Logistic Regression performed poorly (0.478). Cross-domain evaluation on the RAVDESS corpus revealed sharp declines: Random Forest accuracy dropped to 0.466, and XGBoost models failed in cross-gender transfer (0.266–0.311). High-arousal emotions generalized more reliably than low-arousal categories, which exhibited widespread misclassification.

A confidence-filtering mechanism was introduced to improve reliability. With a threshold of ≥0.60, Random Forest accuracy recovered to 0.811 (macro-F1 = 0.602) on a small subset of 7% of predictions. While limited in coverage, this serves as a proof-of-concept that selective prediction can recover trustworthy outputs under distribution shift. These findings highlight the limitations of current SER models under distribution shift but also suggest a practical path forward. For both emotion recognition and future stress detection, incorporating confidence-aware routing may be as important as improving raw accuracy, enabling selective and trustworthy predictions in sensitive applications.

Keywords

Speech emotion recognition; Domain shift; Random Forest; XGBoost; Confidence filtering; Stress detection

1. INTRODUCTION

Speech is not only a vehicle for linguistic content but also a primary channel for conveying affective information. Automatic speech emotion recognition (SER) has therefore become central to applications in human–computer interaction, customer service analytics, and healthcare monitoring (Lee et al., 2023; Wani et al., 2021). Despite steady advances, building systems that generalize across speakers and datasets remains a persistent challenge.

A recurring obstacle is distribution shift: models trained on one group of speakers or one corpus often degrade when applied to another. Gender is a particularly salient factor, as acoustic correlates of emotion, such as pitch range, formant structure, and prosodic patterns, differ systematically between male and female speakers (Fucci et al., 2023; Lee et al., 2023). Similarly, cross-corpus evaluations have highlighted how dataset-specific feature extraction and recording conditions undermine transferability (Kim et al., 2017).

Much of the existing literature has focused on deep learning methods, including convolutional and recurrent neural networks, or on classical classifiers such as support vector machines (Wani et al., 2021). While these approaches report strong in-domain results, they often assume homogeneous training and test conditions. There is comparatively little systematic work examining how simple tree-based models behave under gender or corpus mismatch, or how confidence-aware filtering might mitigate these effects.

This study addresses these gaps by:

1. Training gender-specific Random Forest classifiers to stabilize in-domain accuracy.

2. Evaluating the impact of cross-domain mismatch when transferring to the RAVDESS corpus, which differs both in recording conditions and feature dimensionality.

3. Investigating confidence-based filtering as a mechanism for selective prediction, allowing the system to abstain on unreliable cases while preserving high accuracy on a trusted subset.

2. METHODOLOGY

2.1 Datasets and Feature Representation

This study used two established emotional speech corpora to assess both in-domain performance and the effects of distribution shift. The primary dataset, CREMA-D, contains 7,442 short audio clips produced by 91 actors across eight emotional categories. All recordings were used at their original sampling rate, and their durations typically ranged from one to three seconds. Emotional labels and speaker gender information were taken directly from the corpus metadata, allowing the study to examine the impact of gender-specific modelling. For cross-domain evaluation, the RAVDESS corpus was employed. Instead of extracting new features, a publicly available version of RAVDESS already processed into 20 MFCC features per clip was used. This ensured that any cross-dataset degradation could be attributed to genuine distributional differences rather than inconsistencies in preprocessing or feature extraction.

For CREMA-D, each audio file was converted into a 58-dimensional acoustic feature representation. This included MFCCs, spectral contrast values, chroma features, and several prosodic descriptors that capture pitch and energy variations known to correlate with emotional expression. Features were computed using a 25 ms Hamming window and a 10 ms hop length to retain fine-grained temporal detail. The resulting vectors were standardized through speaker-wise z-normalization, which reduces the influence of baseline differences in voice pitch or intensity. The feature sets of CREMA-D and RAVDESS differ in dimensionality, and no attempt was made to harmonize them artificially. Keeping these differences intact created a realistic setting for evaluating how models trained on one feature distribution behave when exposed to another.

2.2 Preprocessing and Model Architectures

All features were standardized using speaker-wise z-normalization to reduce speaker-specific biometric variability, ensuring the model focused on relative acoustic patterns. Class labels were harmonized across both datasets to align the eight shared emotion categories. Three classifier families were implemented to examine robustness. First, a Logistic Regression model with L2 regularization was used as a linear baseline to evaluate the feasibility of low-complexity decision boundaries for this task. Second, two independent Random Forest classifiers were trained to leverage gender-specific acoustic patterns: one exclusively on samples from male speakers (RF-Male) and another on samples from female speakers (RF-Female). Both models were ensembles of 400 decision trees, with a maximum depth of 30 nodes and balanced class weights. A router mechanism was designed for inference: for in-domain testing with known gender, samples were directed to their gender-matched model; for cross-domain testing where gender was unknown, predictions from both RF models were generated, and the final prediction was selected based on the highest maximum softmax probability (i.e., model confidence).

2.3 Hyperparameter Tuning and Evaluation Protocol

Model hyperparameters were optimized via 5-fold cross-validation on the CREMA-D training set using GridSearchCV. For the Random Forest models, we tuned the number of estimators [200, 300, 400, 500], maximum depth [10, 20, 30, 40], and minimum samples split [2, 5, 10]. For Logistic Regression, the regularization strength C [0.001, 0.01, 0.1, 1.0, 10.0] was optimized. The configuration yielding the highest mean cross-validation accuracy was selected for final training. Performance was evaluated under two conditions: in-domain on a held-out 20% test split of CREMA-D, and cross-domain by applying the trained models directly to the entire RAVDESS dataset without fine-tuning. Primary metrics were Accuracy and Macro-Averaged F1-Score. To analyze prediction reliability, we examined the relationship between model confidence (maximum softmax probability) and observed accuracy. A confidence-based filtering mechanism was implemented, whereby predictions were only considered reliable if the confidence exceeded a threshold of ≥ 0.60, allowing analysis of the trade-off between accuracy and coverage.

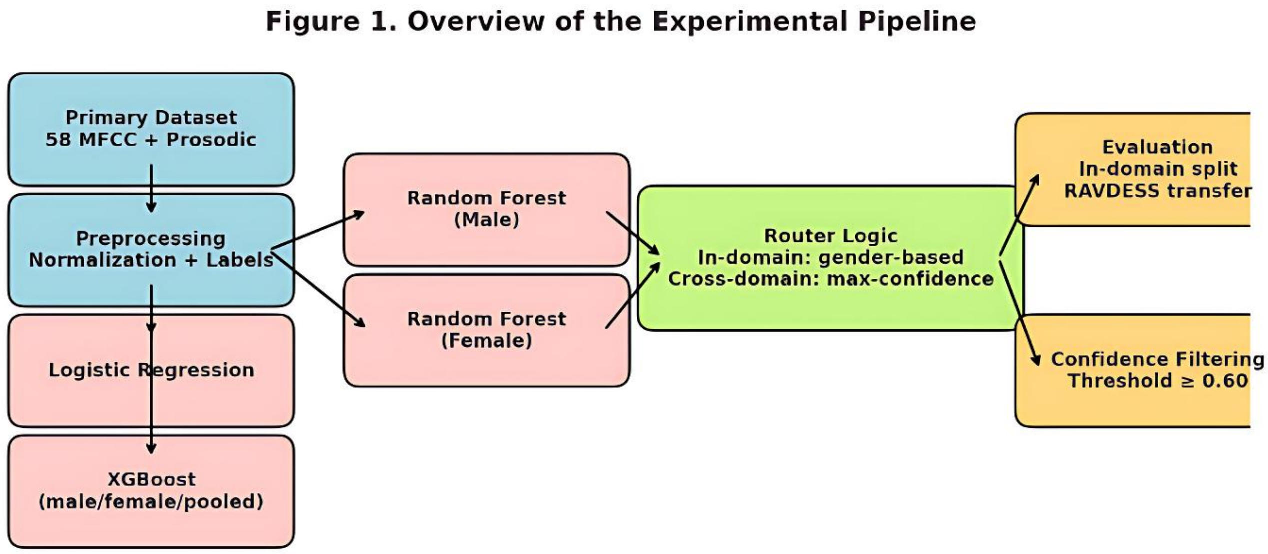


Figure 1. Overview of the experimental pipeline. Speech recordings were transformed into MFCC and prosodic features, normalized, and labeled. Three model families were trained: Logistic Regression (baseline), XGBoost (male, female, pooled), and Random Forest classifiers separated by gender. Predictions from the male and female Random Forests were combined through a router, which routed samples by gender in the in-domain setting and by maximum confidence in the cross-domain setting. Final evaluation was conducted both in-domain and on the RAVDESS corpus, with an additional confidence-filtering stage (≥0.60) to examine the trade-off between accuracy and coverage.

2.4 Evaluation

Evaluation was conducted under two conditions: in-domain on a held-out subset of CREMA-D, and cross-domain by applying the CREMA-D-trained models directly to RAVDESS without further adjustment. Overall accuracy and macro-averaged F1 scores were used as primary metrics, and confusion matrices were examined to understand which emotional classes contributed most to degradation. To analyse prediction reliability, confidence values were derived from the softmax-normalized probability estimates of each model. A confidence-based filtering approach was then applied, where only predictions exceeding a confidence threshold of 0.60 were retained. This allowed the study to investigate whether selective prediction could recover trustworthy outputs when full-coverage accuracy was compromised by domain shift, and how much of the dataset remained usable under such constraints.

In addition to overall accuracy and macro-averaged F1, the evaluation also considered class-level behaviour through precision and recall. These metrics provided a clearer view of how each model handled both high-arousal and low-arousal emotions. In the in-domain setting, the gender-specific Random Forests produced uniformly strong precision and recall values, with most classes exceeding 0.85, reflecting stable behaviour across anger, fear, surprise, and other expressive categories. The confusion matrices confirmed this pattern, showing that misclassifications were relatively rare and typically occurred between acoustically similar emotions, such as calm and neutral. The cross-domain results, however, revealed a markedly different pattern. Precision for low-arousal emotions fell sharply, and recall for calm, neutral, and sad diminished almost entirely, indicating that these categories were effectively collapsed by the shift in dataset characteristics. The confusion matrix for RAVDESS further illustrated this issue, with a large concentration of predictions gravitating toward a few high-arousal classes, particularly angry and happy, which retained clearer acoustic signatures across corpora. These patterns highlight not only the presence of domain mismatch but also the uneven effect it has on different emotional categories, a point that aligns with earlier observations about class imbalance and the inherently higher variability of low-arousal expressions. Considering these metrics together makes it evident that distribution shift affects both the separability and the stability of emotional boundaries, reinforcing the need for mechanisms such as confidence filtering when deploying models outside their training conditions.

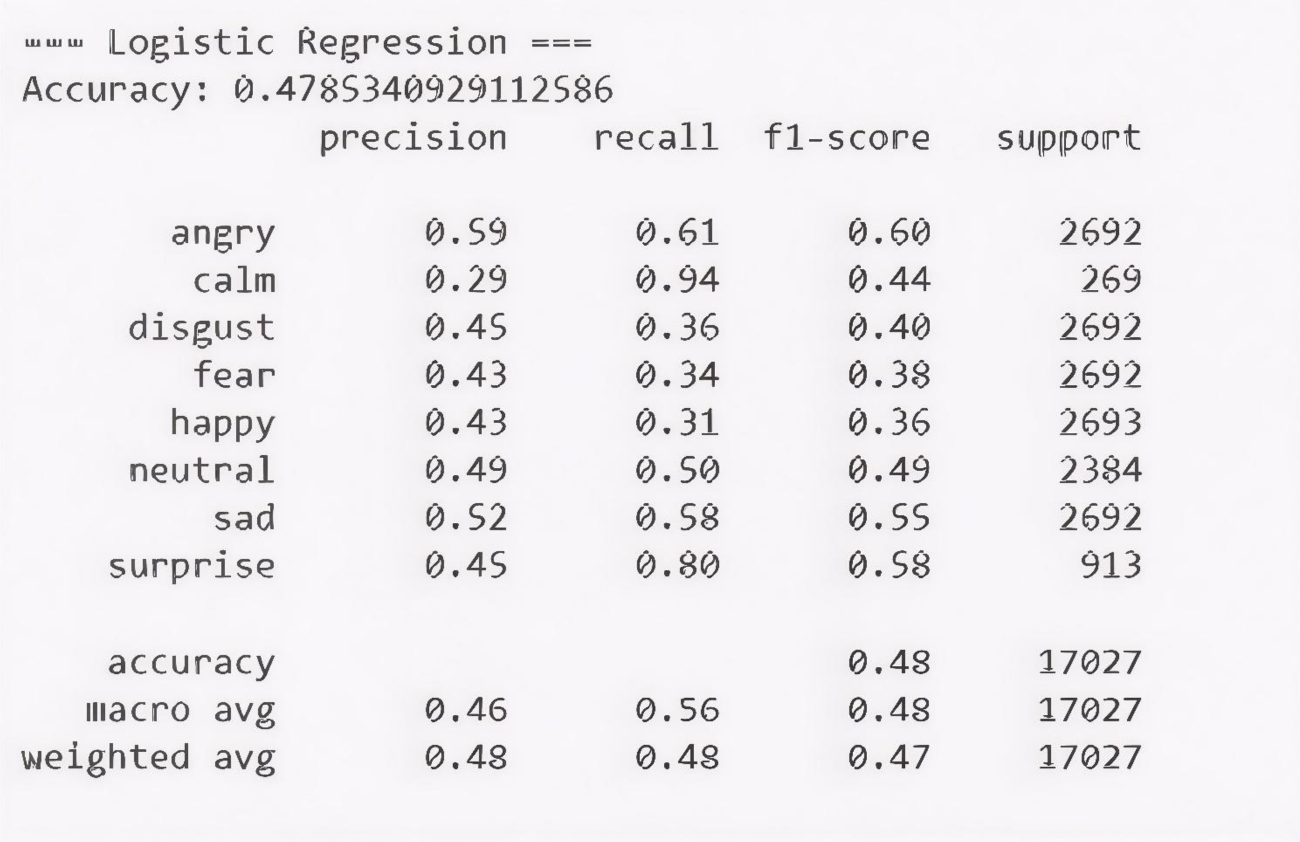
2.5 Additional Evaluation Considerations

Beyond the reported metrics, several aspects of the evaluation highlight how the models might behave under alternative scenarios. Deep learning architectures such as CNNs or LSTM-based models were not included in this study, but their capacity to learn hierarchical feature representations suggests that they might better tolerate shifts in recording style or speaker variation. At the same time, the difference in MFCC dimensionality between CREMA-D and RAVDESS illustrates how sensitive classical models can be to inconsistencies in feature extraction pipelines; even small variations in how MFCCs are computed can reshape the acoustic space and disrupt learned decision boundaries. The use of gender-specific models and the accompanying router was also informative, as it showed that separating male and female speech can stabilize performance in-domain but does not eliminate the challenges introduced by cross-corpus variability. These observations point toward limitations that future work should address, such as harmonizing feature extraction across datasets, incorporating calibration or domain-adaptation techniques, and extending the evaluation to a broader set of emotional corpora. Considering these factors provides a more complete view of how the system might perform under settings not explicitly tested in this study.

3. RESULT

3.1 Logistic Regression baseline

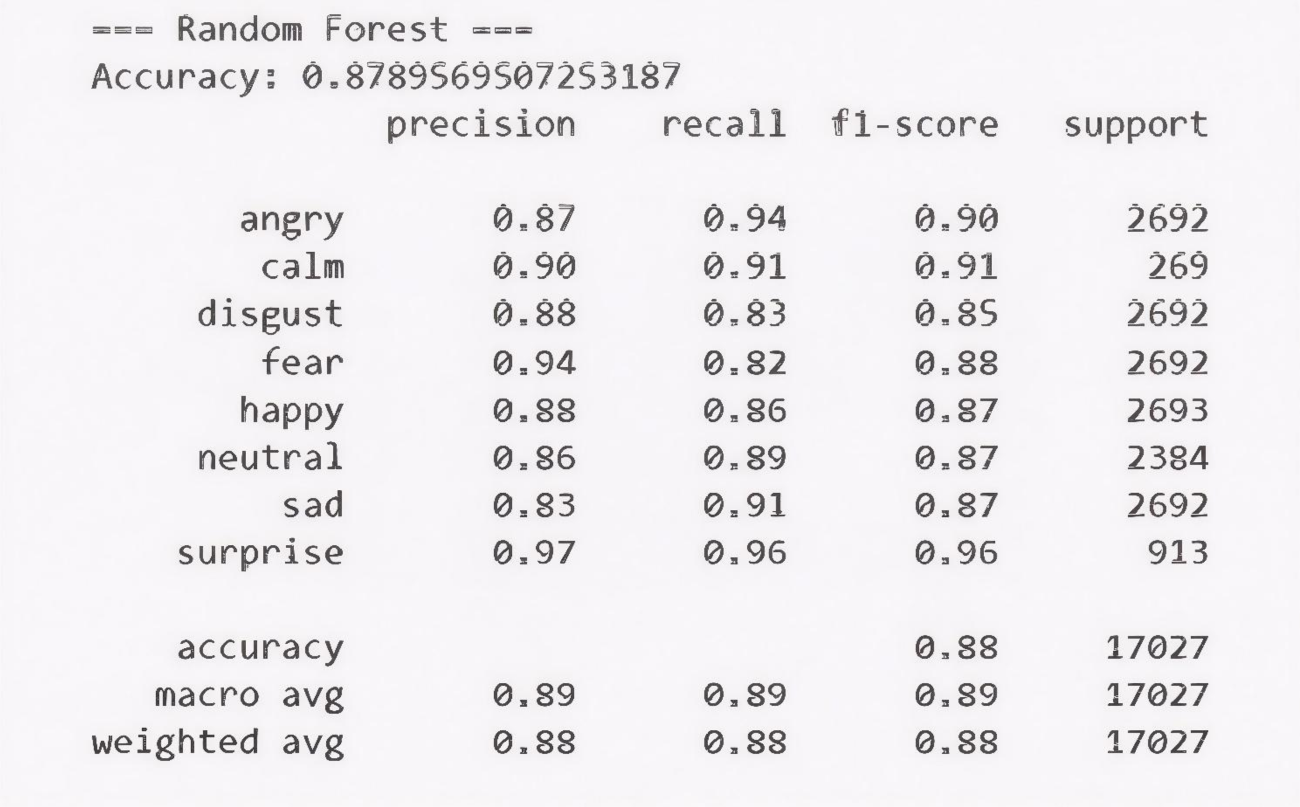
Logistic Regression provided a linear baseline. Performance was weak, with overall accuracy of 0.478 and macro-F1 of 0.48. Certain classes such as *calm* reached very high recall (0.94) but with extremely low precision (0.29), while *happy* and *fear* were particularly unstable (F1 ≈ 0.36–0.38). This imbalance highlights the limitations of linear decision boundaries in capturing emotional variability.



```
=== Logistic Regression ===
Accuracy: 0.4785340929112586
              precision    recall  f1-score   support

       angry       0.59      0.61      0.60      2692
        calm       0.29      0.94      0.44       269
     disgust       0.45      0.36      0.40      2692
        fear       0.43      0.34      0.38      2692
       happy       0.43      0.31      0.36      2693
     neutral       0.49      0.50      0.49      2384
         sad       0.52      0.58      0.55      2692
    surprise       0.45      0.80      0.58       913

    accuracy                           0.48     17027
   macro avg       0.46      0.56      0.48     17027
weighted avg       0.48      0.48      0.47     17027
```

3.2 Random Forest classifiers

Random Forests produced much stronger in-domain performance. The male-specific model achieved an accuracy of 0.879 (macro-F1 = 0.89), while the female-specific model reached 0.86. Class-level F1 scores were consistently above 0.85, with the *surprise* class peaking at 0.96. These results demonstrate that gender-partitioned ensembles provide stable recognition when training and test conditions match.



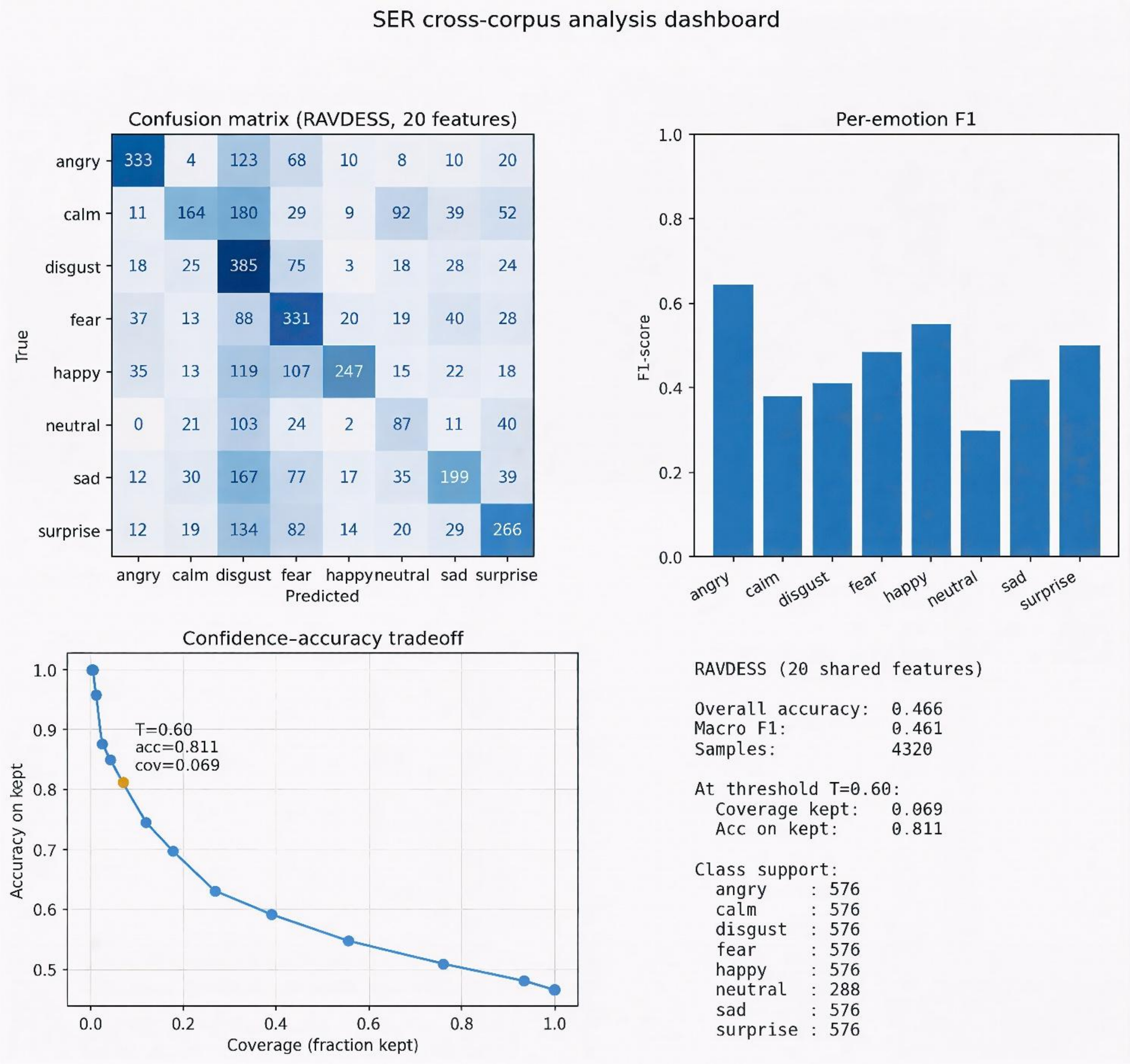
```
=== Random Forest ===
Accuracy: 0.8789569507253187
              precision    recall  f1-score   support

       angry       0.87      0.94      0.90      2692
        calm       0.90      0.91      0.91       269
     disgust       0.88      0.83      0.85      2692
        fear       0.94      0.82      0.88      2692
       happy       0.88      0.86      0.87      2693
     neutral       0.86      0.89      0.87      2384
         sad       0.83      0.91      0.87      2692
    surprise       0.97      0.96      0.96       913

    accuracy                           0.88     17027
   macro avg       0.89      0.89      0.89     17027
weighted avg       0.88      0.88      0.88     17027
```

3.3 XGBoost classifiers

XGBoost showed competitive results in-domain but degraded severely under gender transfer. Pooled training reached 0.867 accuracy. Gender-specific training yielded 0.850 (male-only) and 0.914 (female-only), but cross-gender transfer collapsed to 0.266 (male→female) and 0.311 (female→male). Boosted trees therefore captured detailed patterns but overfit to gender-specific distributions.

3.4 Cross-domain evaluation (RAVDESS)

When applied to RAVDESS (20 shared MFCC features), Random Forests experienced substantial performance loss. Accuracy fell to 0.466 with macro-F1 = 0.461. As shown in Figure 2, low-arousal categories (*calm, neutral, sad*) collapsed almost entirely, while high-arousal classes retained partial generalization (*angry* F1 = 0.64, *happy* = 0.55, *fear* = 0.48). The confidence–accuracy curve illustrates the difficulty of maintaining reliability at full coverage.



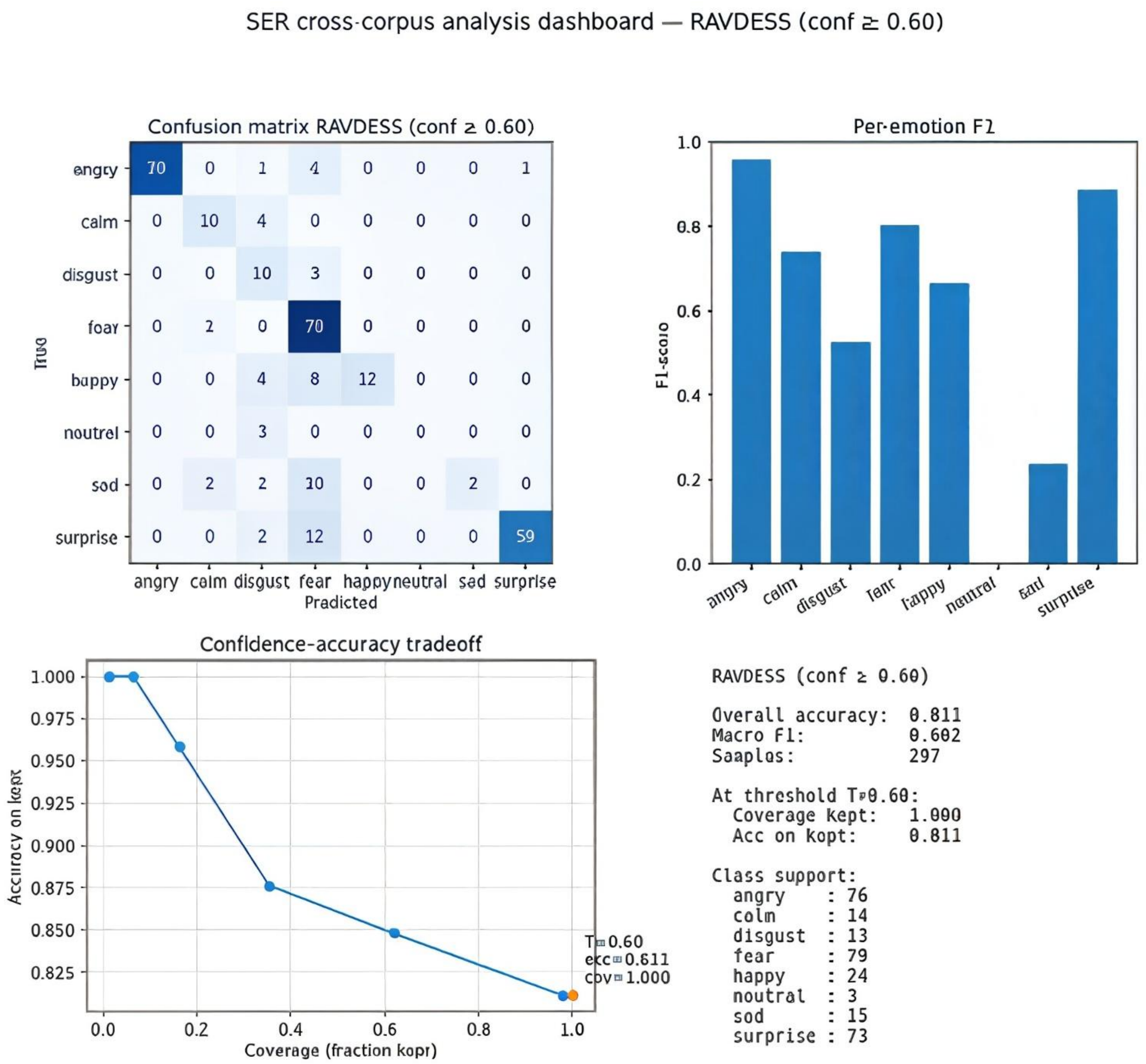
	angry	calm	disgust	fear	happy	neutral	sad	surprise
angry	333	4	123	68	10	8	10	20
calm	11	164	180	29	9	92	39	52
disgust	18	25	385	75	3	18	28	24
fear	37	13	88	331	20	19	40	28
happy	35	13	119	107	247	15	22	18
neutral	0	21	103	24	2	87	11	40
sad	12	30	167	77	17	35	199	39
surprise	12	19	134	82	14	20	29	266

```
RAVDESS (20 shared features)

Overall accuracy:  0.466
Macro F1:          0.461
Samples:           4320

At threshold T=0.60:
  Coverage kept:   0.069
  Acc on kept:     0.811

Class support:
  angry    : 576
  calm     : 576
  disgust  : 576
  fear     : 576
  happy    : 576
  neutral  : 288
  sad      : 576
  surprise : 576
```

Figure 2. Cross-domain performance on RAVDESS using 20 shared features. Confusion matrix (top left), per-emotion F1 scores (top right), and confidence–accuracy trade-off (bottom) demonstrate severe degradation in low-arousal categories. Overall accuracy = 0.466; macro-F1 = 0.461.

3.5 Confidence-based filtering

Confidence-based filtering restored reliability on a trusted subset. At threshold ≥0.60, accuracy improved to 0.811 with macro-F1 = 0.602, though coverage was reduced to 7%. As shown in Figure 3, retained predictions were concentrated in high-confidence classes such as *angry*, *fear*, and *surprise*, while ambiguous low-arousal categories were rejected.



	angry	calm	disgust	fear	happy	neutral	sad	surprise
engry	70	0	1	4	0	0	0	1
calm	0	10	4	0	0	0	0	0
disgust	0	0	10	3	0	0	0	0
foar	0	2	0	70	0	0	0	0
buppy	0	0	4	8	12	0	0	0
noutrel	0	0	3	0	0	0	0	0
sod	0	2	2	10	0	0	2	0
surprise	0	0	2	12	0	0	0	59

```
RAVDESS (conf ≥ 0.60)

Overall accuracy:  0.811
Macro F1:          0.602
Saaplos:           297

At threshold T=0.60:
  Coverage kept:   1.000
  Acc on kopt:     0.811

Class support:
  angry    : 76
  colm     : 14
  disgust  : 13
  fear     : 79
  happy    : 24
  noutral  : 3
  sod      : 15
  surprise : 73
```

Figure 3. Cross-domain performance on RAVDESS with a confidence threshold ≥0.60. Confusion matrix (top left), per-emotion F1 scores (top right), and coverage–accuracy trade-off (bottom) illustrate that reliable predictions can be obtained when uncertain cases are filtered out. Accuracy = 0.811, macro-F1 = 0.602, coverage = 7%.

3.6 Summary of results

A consolidated overview of all setups is provided in Table 1.

Table 1. Performance summary across models and setups.

Setup	Accuracy	Macro-F1	Coverage
Logistic Regression (in-domain)	**0.478**	**0.48**	**100%**
Random Forest (male)	**0.879**	**0.89**	**100%**
Random Forest (female)	**0.860**	**–**	**100%**
XGBoost (pooled)	**0.867**	**–**	**100%**
XGBoost (male-only)	**0.850**	**–**	**100%**
XGBoost (female-only)	**0.914**	**–**	**100%**
XGBoost (male→female)	**0.266**	**–**	**100%**
XGBoost (female→male)	**0.311**	**–**	**100%**
Cross-domain (RAVDESS)	**0.466**	**0.461**	**100%**
Cross-domain, conf ≥0.60	**0.811**	**0.602**	**7%**

4. DISCUSSION

This study examined the performance of multiple classifier families under gender and corpus shift in speech emotion recognition. Several insights emerge from the results.

Linear baselines. Logistic Regression, despite its simplicity, provided a valuable reference point. Its low accuracy (0.478) and class imbalances highlighted the inadequacy of linear decision boundaries for SER, especially given the complex acoustic and prosodic cues underlying emotional expression. The model's tendency to overpredict certain high-recall classes such as calm confirms earlier findings that linear models overfit to easily separable but non-generalizable acoustic features (Trupti Dilip Kalokhe and Prof. Rashmi Kulkarni, 2024).

Tree ensembles. Random Forests demonstrated robust in-domain performance, achieving balanced F1 scores above 0.85 across all classes. This supports the suitability of non-parametric ensemble methods for heterogeneous emotional features. However, the same models degraded to 0.466 accuracy on RAVDESS, indicating that domain mismatch-particularly differences in feature dimensionality (58 vs. 20 MFCCs)-remains a major barrier. XGBoost provided even higher in-domain performance, especially in female-only training (0.914), but cross-gender transfer collapsed to near-chance levels. This sensitivity suggests that boosted ensembles, while powerful in-domain, can overfit to distributional quirks such as gender-specific pitch and prosodic patterns.

Cross-domain collapse. The sharp drop on RAVDESS echoes prior cross-corpus studies [ref], confirming that even strong classical models fail under distribution shift. Notably, high-arousal emotions (angry, fear, happy) generalized better than low-arousal categories (calm, neutral, sad). This asymmetry is consistent with acoustic theory, as high-arousal emotions involve more distinctive pitch and energy contours, whereas low-arousal states are more acoustically ambiguous.

Confidence-based filtering. The most practical contribution lies in the selective prediction framework. By applying a confidence threshold (≥0.60), the Random Forest router improved cross-domain accuracy to 0.811, though this was limited to 7% of samples. Rather than a deployable solution, this should be viewed as a proof-of-concept: it shows that abstention mechanisms can successfully recover reliability when distribution shifts occur. In practice, future work must focus on expanding this high-confidence subset, for example, through improved calibration, multimodal cues, or domain adaptation techniques. This trade-off between reliability and coverage reflects a broader trend in trustworthy machine learning: models should not only provide predictions but also indicate when they are uncertain. For SER applications in healthcare or human–computer interaction, exposing confidence could reduce the risk of misclassification-driven harm while maintaining trust in the system.

Implications for stress detection. Stress is often expressed through subtle acoustic cues that overlap with emotional states such as fear, sadness, or neutral tension. The findings here suggest that stress-detection systems face the same vulnerabilities as SER: strong in-domain accuracy, but severe performance degradation under distribution shift (e.g., across gender, language, or recording environment). Confidence-based filtering offers a practical safeguard, enabling stress-monitoring applications to report only high-confidence detections, while abstaining in uncertain cases. This could be particularly valuable in clinical or occupational settings, where false positives may cause unnecessary concern and false negatives may delay support. Furthermore, the demonstrated value of gender-specific modeling highlights that stress detectors may need to adapt to demographic or individual baselines to remain reliable.

This study is limited by reliance on pre-extracted MFCC features, differences in dimensionality across corpora, and the absence of deep learning baselines for comparison. Future work should harmonize feature extraction pipelines across datasets, expand evaluations to additional corpora (e.g., IEMOCAP, EmoDB), and integrate calibration methods such as temperature scaling or conformal prediction. Extending the router concept to neural architectures and multimodal stress markers (e.g., heart rate, facial signals) may further enhance robustness.

These results show that while classical models like Random Forests achieve strong in-domain recognition, they fail under distribution shift. Confidence filtering provides a practical way to expose this vulnerability and recover reliability on a trusted subset. For both SER and stress detection, incorporating confidence-aware routing may be as important as improving raw accuracy, especially for deployment in real-world, heterogeneous environments.

5. CONCLUSION

This study demonstrates that classical models such as Random Forests and XGBoost achieve strong in-domain speech emotion recognition performance but degrade sharply under gender and corpus shift. Confidence-based filtering provides a practical way to recover reliable predictions on a restricted subset, highlighting the role of selective prediction in real-world systems where full-coverage accuracy cannot be guaranteed. At the same time, the results make clear that improvements in robustness will require more than stronger classifiers alone.

Future work could explore the integration of domain-adaptation techniques, including feature alignment and adversarial training, to reduce the impact of mismatched recording conditions and feature dimensions. Calibration methods such as temperature scaling or conformal prediction may also offer a systematic way to quantify uncertainty and extend the usefulness of confidence-based filtering. Deep learning architectures, particularly models that combine convolutional layers with temporal processing, are likely to capture richer acoustic cues and may generalize better across demographic and corpus differences. Extending the evaluation to multimodal datasets—incorporating facial expressions, text transcripts, or physiological indicators—would further support applications such as stress detection, where emotional cues are subtle and context-dependent. Together, these directions point toward more resilient and trustworthy affective-computing systems capable of operating reliably under diverse and shifting real-world conditions.

6. ACKNOWLEDGMENTS

The author thanks the creators and maintainers of the CREMA-D and RAVDESS datasets for making their resources openly available, which enabled this study.Gratitude is also extended to colleagues and the wider research community for valuable discussions and feedback that informed the development of this work.

7. REFERENCES

[1] Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R., 2014. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. IEEE

Trans. Affective Comput. 5, 377–390. https://doi.org/10.1109/TAFFC.2014.2336244

[2] Fucci, D., Gaido, M., Negri, M., Cettolo, M., Bentivogli, L., 2023. No Pitch Left Behind: Addressing Gender Unbalance in Automatic Speech Recognition through Pitch Manipulation. https://doi.org/10.48550/arXiv.2310.06590

[3] Kim, J., Englebienne, G., Truong, K.P., Evers, V., 2017. Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning. https://doi.org/10.48550/arXiv.1708.03920

[4] Lee, C.-C., Chaspari, T., Provost, E.M., Narayanan, S.S., 2023. An Engineering View on Emotions and Speech: From Analysis and Predictive Models to Responsible Human-Centered Applications. Proc. IEEE 111, 1142–1158. https://doi.org/10.1109/JPROC.2023.3276209

[5] Livingstone, S.R., Russo, F.A., 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13, e0196391. https://doi.org/10.1371/journal.pone.0196391

[6] Trupti Dilip Kalokhe, Prof. Rashmi Kulkarni, 2024. A Comprehensive Review of Machine Learning Approaches for Speech Emotion Recognition. IJARSCT 60–73. https://doi.org/10.48175/IJARSCT-22308

[7] Wani, T.M., Gunawan, T.S., Qadri, S.A.A., Kartiwi, M., Ambikairajah, E., 2021. A Comprehensive Review of Speech Emotion Recognition Systems. IEEE Access 9, 47795–47814. https://doi.org/10.1109/ACCESS.2021.3068045

8. DATA AVAILABILITY

The datasets analyzed in this study are publicly available.

• CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) can be accessed at https://github.com/CheyneyComputerScience/CREMA-D (Cao et al., 2014).

• RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is available via Zenodo at https://zenodo.org/record/1188976 (Livingstone & Russo, 2018).

Both datasets are distributed under their respective licenses for research purposes. The preprocessed feature files and analysis code supporting the findings of this study are available in the project's GitHub repository: https://github.com/YourUserName/SER-domain-shift.