ARMEN

ADVERSARIALLY RESILIENT NETWORK FOR EMOTION RECOGNITION FROM MUSIC

Paper ID : 158

Authors: Renu Akter Suity, Niamul Hassan Samin, Bijon Mallik, Ahmed Shafkat, Md. Mahbubur Rahman, Kohinur Parvin

Affiliation: Bangladesh University of Business and Technology (BUBT), University of Wyoming, Netrokona University

Outline

- Introduction
- Motivation
- Literature Review
- Methodology
- Result Analysis
- Limitations
- Future Work
- Conclusion
- Reference

“Imagine your favorite **music** app knows exactly how you feel — it plays calm tunes when you’re stressed and energetic beats when you’re motivated**.**
But what if a tiny, invisible glitch in the data suddenly makes it think you’re sad when you’re actually happy?
That’s not just a bug — **that’s an adversarial attack.**
Our research, ARMEN, is designed to stop that.”

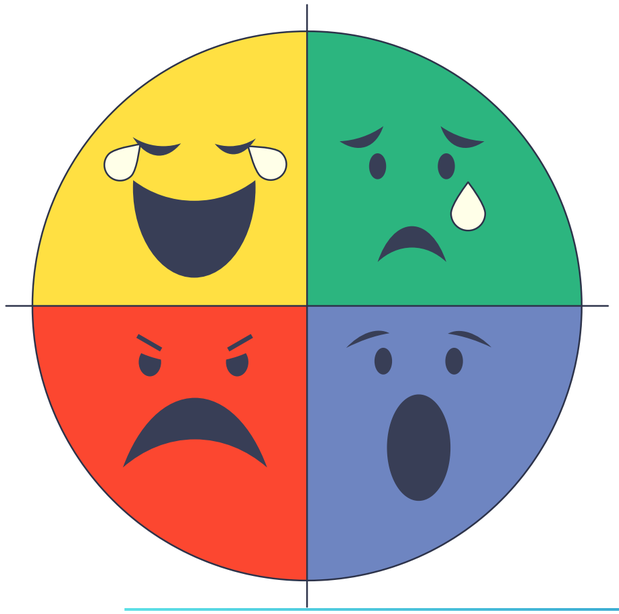
To be precise,

an **adversarial attack** is when tiny and carefully designed changes are added to input data **(like audio features, images, or text)** to trick a machine learning model into making the wrong prediction — even though the changes are almost invisible to humans.

Presenting our Solution!

ARMEN

ARMEN is a robust and interpretable machine learning framework designed to withstand adversarial perturbations while maintaining high classification accuracy on emotions in music — **like happy, sad, energetic, or calm.**



06

Music is deeply connected to human emotions, and modern streaming platforms rely on emotion recognition models to personalize user experiences. However, these models are often vulnerable to adversarial attacks (which we talked earlier) -can reduce system trust and reliability.

-This **motivated** the development of **ARMEN**

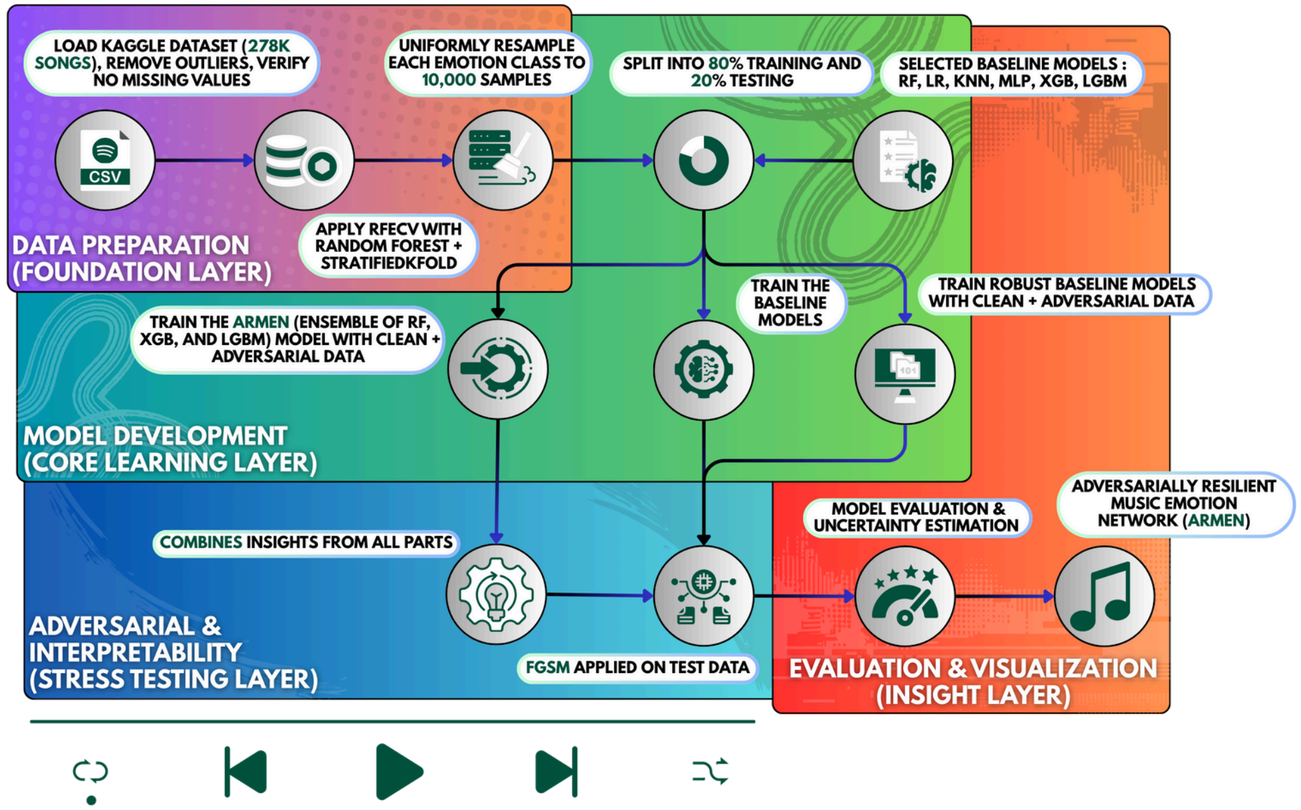
07

Literature Review

Study	Approach	Limitation
Revathy et al. (2023) – LyEmoBERT	BERT for lyric emotion	No audio features
Agrawal et al. (2021) – XLNet	Transformer-based MER	Lyrics only
Hoedt et al. (2022) – MP Defense	Adversarial recommender defense	High cost, not for audio
Huang et al. (2021) – GAN for MER	Deep GAN architecture	No robustness evaluation
ARMEN (This Work)	Ensemble + Adversarial Defense	Robust + Interpretable

Methodology

Overview	Proponents
ARMEN is a stacked **ensemble model** that integrates: • **Random Forest (RF)** • **XGBoost (XGB)** • **LightGBM (LGBM)** These three models work together, and their outputs are combined for the final decision.	To make the model robust, ARMEN includes two defense mechanisms: • **FGSM Adversarial Training** → trains on both normal and perturbed data, so it learns to resist attacks. • **Monte Carlo Dropout** → estimates uncertainty, helping the model measure how confident it is in each prediction.



Moodify

- **278,000 Spotify songs**
- **Features: danceability, energy, valence, tempo, etc.**
- **4 Emotion Labels: Happy, Sad, Energetic, Calm**
- **Balanced subset: 40,000 samples (10K per class)**

ARMEN was trained on the **Moodify dataset** and achieved:

- **96% accuracy on clean data**
- **94% accuracy even under FGSM attacks**

Along with it, **six** popular machine learning models were also trained on both clean and FGSM-perturbed (adversarial) data.
These models include: **Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), Multi-Layer Perceptron (MLP), XGBoost, and LightGBM**.

FGSM (Fast Gradient Sign Method) is a quick method to create adversarial examples by adding a small, targeted perturbation — in the direction of the model's loss gradient — that makes the model misclassify the input while the change remains imperceptible to humans.

Infographics (1)

of six popular machine learning models trained on the Moodify Dataset (Clean and Adversarial Data)

BASELINE MODEL ACCURACY ACROSS CLASSIFIERS

Model	Accuracy
Random Forest	84.3%
Logistic Regression	80.8%
KNN	76.8%
MLP	82.8%
XGBoost	85.8%
LightGBM	**86.3%**

MODEL ACCURACY ON CLEAN AND ADVERSARIAL DATA BEFORE VS. AFTER ADVERSARIAL TRAINING

Model	FGSM Accuracy	Clean Accuracy
Random Forest (Baseline)	62.5%	84.3%
Random Forest (Robust)	80.3%	82.7%
Logistic Regression (Baseline)	57.8%	80.7%
Logistic Regression (Robust)	76.1%	79.4%
KNN (Baseline)	52.4%	76.8%
KNN (Robust)	70.5%	74.6%
MLP (Baseline)	60.9%	82.8%
MLP (Robust)	78.6%	81.1%
XGBoost (Baseline)	64.2%	85.8%
XGBoost (Robust)	82.5%	84.1%
LightGBM (Baseline)	65.0%	**86.3%**
LightGBM (Robust)	**83.4%**	84.9%

Infographics (2)

ARMEN trained on the Moodify Dataset (Clean and Adversarial Data)

ARMEN Classification Report on Clean Data

Emotion	Precision	Recall	F1-score	Support
Happy	0.96	0.95	0.96	250
Sad	0.91	0.93	0.92	250
Energetic	0.95	0.96	0.95	250
Calm	1.00	0.99	0.99	250
Accuracy	0.96			

ARMEN Classification Report on Adversarial Data After Adversarial Training

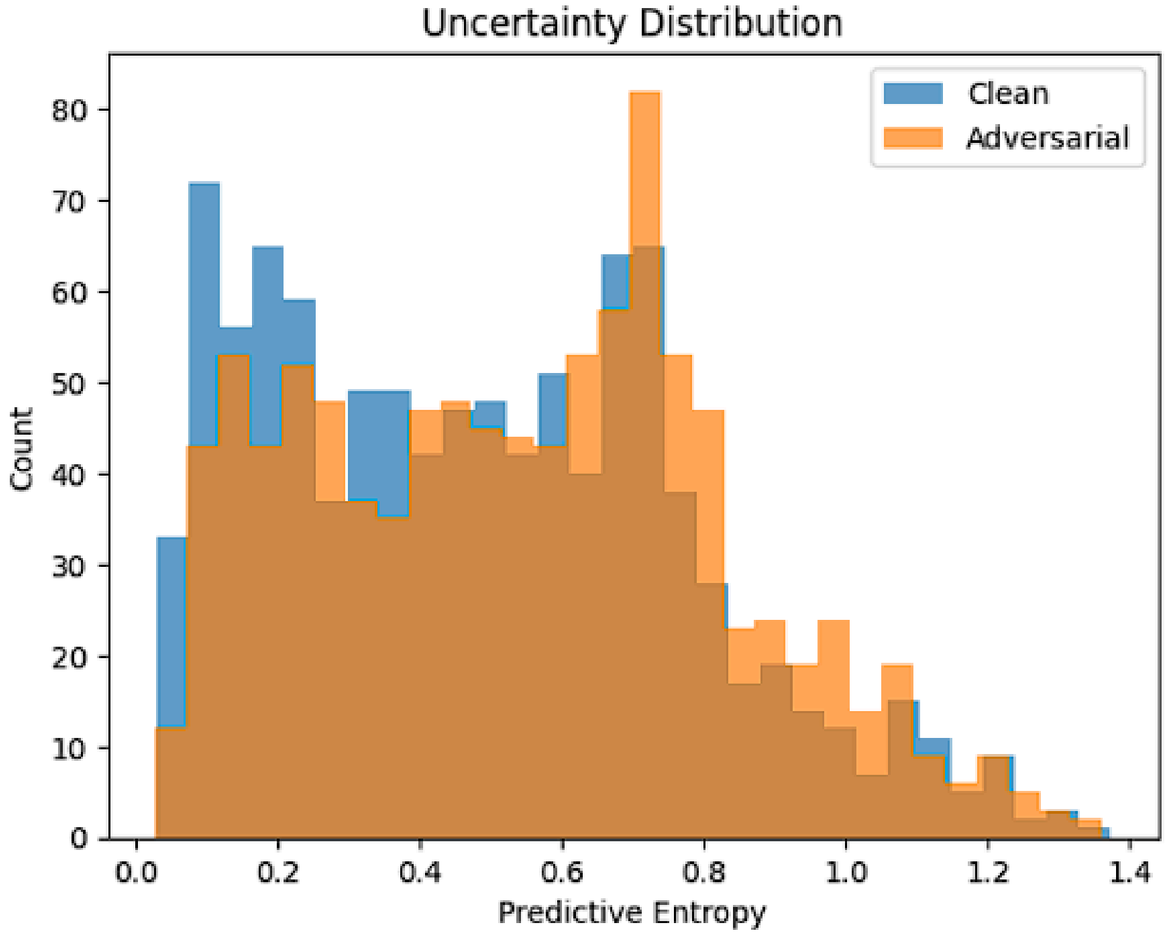
Emotion	Precision	Recall	F1-score	Support
Happy	0.93	0.97	0.95	250
Sad	0.90	0.89	0.89	250
Energetic	0.94	0.91	0.92	250
Calm	1.00	0.99	0.99	250
Accuracy	0.94			

Uncertainty Estimation

Monte Carlo Dropout

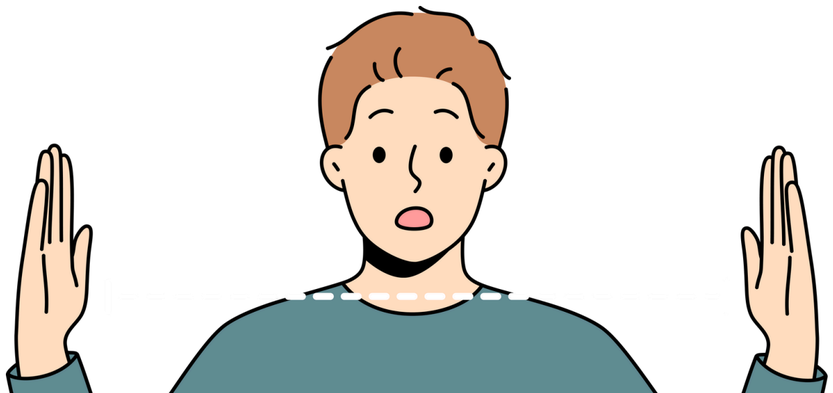
Monte Carlo Dropout helps **ARMEN** estimate how confident it is in each prediction.

By running the same input multiple times with **random neuron drops**, we can detect when the model becomes uncertain — especially useful under adversarial conditions.

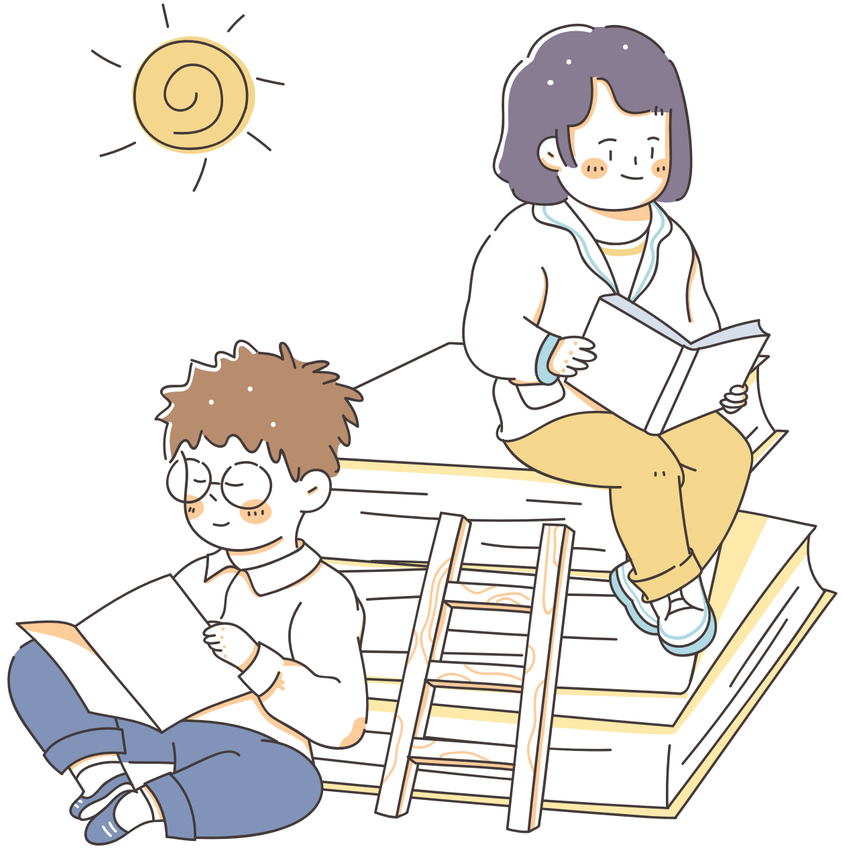


Limitations

- This study focuses only on FGSM attacks — stronger methods like PGD or CW were not tested.
- We used a subset (40,000 samples) of the full 278K Moodify dataset due to computational limits.
- Evaluations were performed on structured audio features only — lyrics and multimodal data were not included.
- Real-time or streaming adaptation was not implemented in this phase.



Future Work



- Apply stronger attacks (PGD, CW) for future robustness testing.
- Explore multimodal integration (lyrics, EEG, and metadata).
- Extend ARMEN to multi-language and cross-genre datasets.

Conclusion

ARMEN — an ensemble-based resilient network — combines **Random Forest, XGBoost,** and **LightGBM** with **FGSM adversarial training** and **Monte Carlo Dropout** to enhance robustness in music emotion recognition. It achieves **96%** accuracy on clean data and **94%** under adversarial attacks, proving strong resistance and interpretability. This work establishes a foundation for secure, reliable, and emotion-aware music recommendation systems.

Reference

[1] P. Babu, V. Singh, and A. Gupta, "Emotion aware music recommendation system using contextual lstm networks," *arXiv preprint arXiv:2311.10796*, 2023.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2015.

[3] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*, 2016, pp. 1050–1059.

[4] A. Werner, "Organizing music, organizing gender: algorithmic culture and spotify recommendations," *Popular Communication*, vol. 18, no. 1, pp. 78–90, 2020.

[5] V. Revathy, A. S. Pillai, and F. Daneshfar, "Lyemobert: Classification of lyrics' emotion and recommendation using a pre-trained model," *Procedia Computer Science*, vol. 218, pp. 1196–1208, 2023.

[6] Y. Agrawal, R. G. R. Shanker, and V. Alluri, "Transformer-based approach towards music emotion recognition from lyrics," in *European conference on information retrieval*. Springer, 2021, pp. 167–175.

[7] E. Shakirova, "Collaborative filtering for music recommender system," in *2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, 2017, pp. 548–550.

[8] S. Wang, C. Xu, A. S. Ding, and Z. Tang, "A novel emotion-aware hybrid music recommendation method using deep neural network," *Electronics*, vol. 10, no. 15, p. 1769, 2021.

[9] S. Joshi, T. Jain, and N. Nair, "Emotion based music recommendation system using lstm-cnn architecture," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*. IEEE, 2021, pp. 01–06.

10] S. Shitole, "Music recommendation system on spotify using deep learning," 2024.

11] T. M. Al-Hasan, A. N. Sayed, F. Bensaali, Y. Himeur, I. Varlamis, and G. Dimitrakopoulos, "From traditional recommender systems to gpt-based chatbots: A survey of recent developments and future directions," *Big Data and Cognitive Computing*, vol. 8, no. 4, p. 36, 2024.

[12] A. B. Melchiorre, D. Penz, C. Ganhör, O. Lesota, V. Fragoso, F. Fritzl, E. Parada-Cabaleiro, F. Schubert, and M. Schedl, "Emotion-aware music tower blocks (emomtb): an intelligent audiovisual interface for music discovery and recommendation," *International Journal of Multimedia Information Retrieval*, vol. 12, no. 1, p. 13, 2023.

[13] M. A. S. Siddique, M. I. Sarker, R. Ghosh, and K. Gosh, "Toxicity classification on music lyrics using machine learning algorithms," in *2021 24th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2021, pp. 1–5.

[14] H. Joung and K. Lee, "Music auto-tagging with robust music representation learned via domain adversarial training," 2024. [Online]. Available: https://arxiv.org/abs/2401.15323

[15] M. Park and K. Lee, "Exploiting negative preference in content-based music recommendation with contrastive learning," in *Proceedings of the 16th ACM Conference on Recommender Systems*, ser. RecSys '22. ACM, Sep. 2022, p. 229–236. [Online]. Available: http://dx.doi.org/10.1145/3523227.3546768

[16] X. Jin, W. Zhou, J. Wang, D. Xu, and Y. Zheng, "An order-complexity aesthetic assessment model for aesthetic-aware music recommendation," 2024. [Online]. Available: https://arxiv.org/abs/2402.08300

[17] O. Lesota, J. Geiger, M. Walder, D. Kowald, and M. Schedl, "Oh, behave! country representation dynamics created by feedback loops in music recommender systems," in *18th ACM Conference on Recommender Systems*, ser. RecSys '24. ACM, Oct. 2024, p. 1022–1027. [Online]. Available: http://dx.doi.org/10.1145/3640457.3688187

[18] S. Hoedt, W. De Neve, and K. Demuynck, "Defending music recommender systems against hubness-based adversarial attacks," *arXiv preprint arXiv:2205.12032*, 2022.

[19] I.-S. Huang, Y.-H. Lu, M. Shafiq, A. Ali Laghari, and R. Yadav, "A generative adversarial network model based on intelligent data analytics for music emotion recognition under iot," *Mobile information systems*, vol. 2021, no. 1, p. 3561829, 2021.

[20] X. Qiu, "Improving robustness in emotion recognition via adversarial training," in *Fourth International Conference on Signal Processing and Machine Learning (CONF-SPML 2024)*, vol. 13077. SPIE, 2024, pp. 119–126.

[21] K. Hoedt, A. Flexer, and G. Widmer, "Defending a music recommender against hubness-based adversarial attacks," 2022. [Online]. Available: https://zenodo.org/record/6573391

[22] "278k Emotion Labeled Spotify Songs — kaggle.com," https://www.kaggle.com/datasets/abdullahorzan/moodify-dataset?select=278k_song_labelled.csv, [Accessed 24-05-2025].

[23] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *Advances in neural information processing systems*, 2017, pp. 3146–3154.

[26] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

Thank you!

Do you have any questions?