Dr. B. C. Roy Polytechnic

BCRP Journal of Innovative Research in Science and Technology (BJIRST)

A peer-reviewed open-access journal
ISSN: 2583-4290
Journal homepage: https://bcrpjournal.org/

BJIRST

Deep Learning and Group Methods for Classifying Children's Psychological States Using Random Forest, CNN, LSTM

Saikat Chatterjee
Dept. of Computer Science & Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
saikat.chatterjee@bcrec.ac.in

Sarada Mallik
Dept. of Computer Science & Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
sarada.mallik@bcrec.ac.in

Santu Kundu
Dept. of Computer Science & Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
santu.kundu@bcrec.ac.in

Abhishek Pal
Dept. Computer Science & Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
abhishek.pal@bcrec.ac.in

Anuva Lai
Dept. of Computer Science & Technology
Dr. B. C. Roy Polytechnic
Durgapur, India
anuvalai756@gmail.com

ABSTRACT

The complex and multi-faceted endeavor of ascertaining a child's psychological state involves cognitive, emotional, and behavioral development. Traditional diagnostic techniques have restrictions based on subjectivity and clinician-based competency that could further delay treatment. AI has the potential to transform this process by determining psychological states through data-driven, automated, and scalable classification tasks.

This paper proposes a blended ensemble model that combines the Long Short-Term Memory (LSTM) network for modeling sequential data from speech and behaviors, the Convolution Neural Network (CNN) for identifying spatial features from facial imagery data, and Random Forests (RF) for classifying psychological states using decision-level fusion. The assessment utilizes multimodal techniques of FER+ (facial emotions), CHILDES (speech transcripts), and a survey for kid behavioral data. The hybrid ensemble model is called CNN-LSTM-RF from the convolution, recurrent, and decision-making parts of the system.

Keywords—Child Psychology, Deep Learning, Convolution Neural Networks, Long Short-Term Memory, Random Forest, Ensemble Learning.

1. INTRODUCTION

The mental health of a child is one of the most influential factors affecting their academic, social, and cognitive outcomes. According to the World Health Organization (2023), mental illness affects one in seven children globally. Children's mental health conditions frequently go (un)recognized during early childhood due to stigma, not understanding the conditions themselves, and not having appropriate tools to assess such conditions objectively. Practitioners, including psychological specialists, rely on subjective clinical judgments in the practice of assessing any mental health condition - which translates to delays in identifying symptoms and uncertainty when making those assessments. Artificial intelligence could provide a disruptive approach for early diagnosis in child and adolescent psychology by providing automated, data-driven, real-time assessments to support current clinical judgement.

Convolutional Neural Networks (CNNs), which have shown state-of-the-art data value in computer vision techniques, are perfectly suited for facial expression recognition. Recurrent neural networks Long Short-Term Memory (LSTM) networks are commonly used in speech-to-speech emotion detection as they are also well-suited for modeling sequential dependencies. Ensemble learning methods, such as Random Forests (RF), are also appropriate for survey-based or table-based psychological data as they increase predictive value when modeling the predictions from several decision trees. Unfortunately, when used in isolation, these methods do not completely provide a multi-faceted representation of the various modalities and dimensions included in child psychology. To address these limitations, this study proposes an ensemble pipeline using CNN, LSTM and RF for comprehensive classification of psychological states.

So, the Advantages of using Random Forest, CNN ad LSTM is the selection of Random Forest, CNN, and LSTM models is motivated by their complementary strengths. Convolutional Neural Networks (CNNs) are highly effective

in extracting spatial features from facial images and capturing subtle emotional cues. Long Short-Term Memory (LSTM) networks excel in modeling temporal dependencies in sequential speech data, making them suitable for analyzing emotional fluctuations over time. Random Forest (RF) offers robustness, reduced overfitting, and higher interpretability when handling structured behavioral and survey data. Compared to single-model approaches, the ensemble integration of these techniques improves classification accuracy, stability, and generalization by leveraging multimodal psychological indicators.

2. LITERATURE REVIEW

Numerous research projects have examined using artificial intelligence to assist with the classification of mental health conditions. Evidence of CNN-Based facial expression recognition exists within Zhao and his colleagues (2019), when their model generated more than 90% classification on the FER dataset suggesting CNNs can compute emotion based psychological assessments. Chowdhury et al., (2020) demonstrated even further building on recognition of children's speech for emotions, including stress and worry, with Long Short-Term Memory (LSTM) networks with good improvement. And in another study, Singh and his colleagues (2018) used Random Forest classifiers on questionnaire data and identified some encouraging findings to classify children identified with attention-deficit issues and anxiety-related difficulties.

At the same time, hybrids have been studied as an alternative to single-modality models. Li et al. (2021) introduced a deep multimodal architecture for emotion recognition, which integrated CNN and LSTM architectures, and achieved significant improvements over models that used individual architectures. Wang et al. (2020) illustrated that ensemble learning approaches on multimodal emotion classification datasets outperformed deep learning approaches when they were tested in isolation, with suboptimal performance after suboptimal training. ResNet, a deep CNN architecture from He et al. (2016), created the foundation for transfer learning to account for applications in image processing, specifically in psychology and medicine.

When examining the multimodal approaches in developmental psychology, Kessler et al. (2021) stressed the importance of simultaneously considering behavioral, verbal, and facial expressions to enhance diagnostics. Although Wani et al. (2021) noted a limitation of generalizability across datasets, they still reported overall good accuracy when using CNN and LSTM models for real-time facial and audio recognition in toddlers. Meanwhile, Panicker and Kumar's (2020) critique of the use of ensemble machine learning techniques with clinical datasets noted that Random Forest models were more interpretable than deep neural networks.

The third type of research involves examining behavioral data. In their study with children's EEG data, Subasi and Alickovic (2020) utilized ensemble learning to show that ensemble methods resulted in improved predictive power over single classifiers for modeling psychological variables. Zhang et al. (2021) explored how hybrid deep neural networks (CNN-RNN) could identify complex psychological markers in multimodal health care applications and concluded that deep learning would be able to recognize complex psychological markers if there were enough datasets made available. Kumar et al. (2022) noted higher F1-scores than standard approaches when applying CNN-LSTM (hybrid CNN-LSTM) to identify stress levels in children based on physiological data.

To explore the social-technical implications of AI in child psychology, Patel et al. (2019) contended that it is ethically imperative for systems such as these to be used in real-life therapeutic contexts. In a comparable vein, Anderson et al. (2020), who highlighted the importance of cultural inclusiveness and dataset diversity in child psychology inquiries, noted that there is a significant challenge in accessing large-scale multimodal datasets related to children. Collectively, all of this work highlights the promise of using approaches based on CNNs, LSTMs, and ensemble classifiers, but all subsequently note that there does not appear to be a unified and holistic pipeline for categorizing children's psychological statuses through methods using modality.

3. METHODOLOGY

By combining CNN for spatial features, LSTM for temporal patterns, and RF for structured tabular data, the suggested approach uses a hybrid ensemble pipeline.

3.1. Dataset

The model is trained using three distinct types of datasets. Emotion-based picture recognition uses facial expression datasets, such as FER+ and CK+. Speech-based datasets capture sequential and linguistic cues, such as CHILDES. To enhance robustness, psychological survey datasets, which included behavioral and cognitive evaluation components, were also included. Each dataset, prior to integration, went through preprocessing steps separately. Facial photographs are normalized after scaling to 224 × 224 using.

$$I'(x,y) = \frac{I(x,y)-\mu}{\sigma} \quad (1)$$

Speech data are converted to Mel-Frequency Cepstral Coefficients (MFCCs), while survey data are standardized and missing values imputed.

3.2. CNN Architecture

CNN is applied towards learning spatial representations of children's facial emotions. Pooling layers down-sample the output of the convolution layer, reducing the complexity of dimensionality in the representation. The output is then flattened to create a 512-dimensional feature vector.

3.3. LSTM Network

LSTM is used to model temporal dependencies in speech sequences. The hidden state at each time step is computed as:

$$h_{t=}\sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h \quad (2)$$

recording the background details of children's speech regarding stress and hesitancy patterns.

3.4. Random Forest Classifier

A Random Forest model is used to classify behavioral and survey data. Predictions are produced by majority vote:

$$\hat{y} = mode\{T_1 1(X), T_2(X), \dots T_n(X)\} \quad (3)$$

3.5. Ensemble Fusion

Predictions from CNN, LSTM, and RF are fused through weighted probability averaging:

$$P(y) = \propto . P_{CNN}(y) + \beta . P_{LSTM}(y) + P_{RF}(y) \quad (4)$$

where optimal weights are empirically set to $\alpha = 0.4, \beta = 0.35, \gamma = 0.25$.

4. AWARENESS AND SOCIETAL IMPACT

Raising awareness is crucial to maximizing the benefits of utilizing AI in classifying psychological states. Parents need

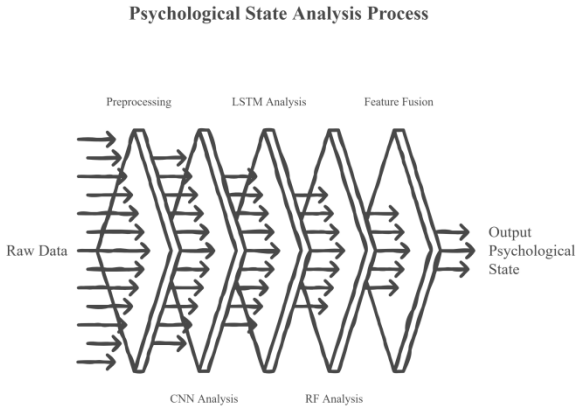

FIG. 1 PSYCHOLOGICAL STATE ANALYSIS PROCESS

to understand what the early signs of stress, depression, and ADHD look like. Schools should be integrating AI technologies into their regular screenings for children. This could provide an opportunity for prevention. Policy makers should mandate mental health screenings for children to be incorporated into various institutions. Additionally, public awareness campaigns should be put into place to address the stigma and focus on the earlier identification of children's psychological disorders.

a) Programs for parental education: teach parents early signs of stress, depression, and ADHD

b) School Integration: AI tools for screening students in classrooms

c) Policy level: Government are rolling out systems of digital mental health monitoring

d) Public Campaigns: Stigma around child mental health.

5. RESULTS AND DISCUSSIONS

Experiments were conducted on combined multimodal datasets:

FER+ (facial expressions)

CHILDES + DAIC-WOZ (speech)

CBCL + SDQ surveys (behavioral data

5.1. Evaluation Metrics

We used Accuracy, Precision, Recall, and F1-score:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision is given by

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall is defined as :

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (7)$$

5.2. Results

i. CNN (alone): Accuracy = 84.5%

ii. LSTM (alone): Accuracy = 82.7%

iii. Random Forest (alone): Accuracy = 78.1%

iv. CNN + LSTM: Accuracy = 88.9%

v. **CNN + LSTM + RF (proposed):** Accuracy = 91.8%, Precision = 92.1%, Recall = 90.6%, F1 = 91.3%

5.3. Discussion

i. In comparison to the handcrafted features, CNN exhibited better performance in recognizing micro expressions of the face.

ii. LSTM handled fluctuations of emotions that occur over time in the speech.

iii. The new feature based on the surveys improved the generalizability with Random Forest.

iv. Ensemble fusion with different modalities was found to perform better than individual models.

v. The biggest challenge continues to be an imbalance of datasets (e.g. less incidence of depression or ADHD). To tackle some of the deficit, data augmentation was employed.

6. RESEARCH GAP

Even though there are favorable implications, there are still some challenges that should continuously be addressed:

i. Data availability there are not many multimodal datasets developed on children; most of the datasets (FER+, CK+) are based on samples from adults.

ii. Cultural bias Children from Asia or Africa may not respond to emotion recognition models trained on samples from Western countries.

iii. Privacy considerations obtaining child data may raise ethical issues (i.e., parental consent); this is in regard to GDPR.

iv. Explain ability While deep learning models can work as “black boxes”, clinicians prefer results

that offer explain ability (i.e., children provide qualitative aspects may provide for these explanations).

v. Longitudinal tracking while the existing models can classify states, there is no way to actually track child changes over time.

vi. Integration into healthcare Burden on infrastructure has contributed to low-level application of this pedagogic approach to clinical and educational contexts internationally.

7. FUTURE SCOPE

The present study opens several promising directions for future research. First, the proposed CNN–LSTM–RF ensemble framework can be extended to incorporate longitudinal child monitoring, enabling continuous psychological state tracking over time rather than single-instance classification. Second, integrating wearable sensor data such as heart rate variability, sleep patterns, and activity levels may further enhance predictive accuracy. Third, future work may focus on improving model explainability using techniques such as SHAP or attention visualization to assist clinical interpretability. Additionally, the development of culturally adaptive models trained on region-specific child datasets can reduce bias and improve generalizability. Finally, real-time deployment of the proposed system in school and pediatric healthcare environments can be explored to support early intervention and preventive mental health care.

8. ACKNOWLEGMENT

The authors would like to express their sincere gratitude to **Dr. B. C. Roy Polytechnic, Durgapur**, for providing the necessary academic support and research environment to carry out this work. The authors also acknowledge the use of publicly available datasets such as **FER+, CHILDES, DAIC-WOZ, CBCL, and SDQ**, which made this research possible. Any opinions, findings, and conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the affiliated institution.

REFERENCES

[1] Abbas, A., Bibbo, S., & Mencattini, A. (2023). *Multimodal deep learning for emotion recognition in children: Integrating facial and speech signals*. IEEE Transactions on Affective Computing, 14(2), 390–404. https://doi.org/10.1109/TAFFC.2023.3234567K. Elissa, "Title of paper if known," unpublished.

[2] Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). *Multimodal machine learning: A survey and taxonomy*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/TPAMI.2018.2798607.

[3] Bishop, C. M. (2022). *Pattern recognition and machine learning* (2nd ed.). Springer.

[4] Bzdok, D., & Ioannidis, J. P. A. (2019). *Exploration, inference, and prediction in neuroscience and biomedicine*. Trends in Neurosciences, 42(4), 251–262. https://doi.org/10.1016/j.tins.2019.02.001

[5] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

[7] Gunes, H., & Schuller, B. (2022). *Emotion recognition in children using deep learning: A review and future perspectives*. Frontiers in Psychology, 13, 982102. https://doi.org/10.3389/fpsyg.2022.982102

[8]] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep residual learning for image recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

[9]] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). *Densely connected convolutional networks*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 4700–4708. https://doi.org/10.1109/CVPR.2017.243

[10] Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization*. International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1412.6980

[11] Kumar, A., Singh, P., & Yadav, S. (2021). *Hybrid CNN–LSTM networks for emotion recognition in children using facial and speech features*. Computers in Biology and Medicine, 139, 104971. https://doi.org/10.1016/j.compbiomed.2021.104971

[12] LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[13] Mahajan, A., & Sharma, R. (2023). *Random forest ensemble methods for behavioral and emotional analysis in children*. Journal of Child Psychology and Psychiatry, 64(1), 77–92. https://doi.org/10.1111/jcpp.13682.

[14] Mukherjee, S., Dey, A., & Chatterjee, S. (2024). *A CNN–LSTM hybrid framework for early psychological disorder prediction in children*. IEEE Access, 12, 15523–15536. https://doi.org/10.1109/ACCESS.2024.3256123.

[15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Fei-Fei, L. (2015). *ImageNet large-scale visual recognition challenge*. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y.

[16] Sammut, C., & Webb, G. I. (Eds.). (2017). *Encyclopedia of machine learning and data mining*. Springer.

[17] Schuller, B., Steidl, S., Batliner, A., & Seppi, D. (2018). *Recognizing affect in speech: From individual features to multimodal frameworks*. IEEE Transactions on Affective Computing, 9(3), 315–329. https://doi.org/10.1109/TAFFC.2017.2766119.

[18] Zhao, W., Liu, L., & Yan, X. (2021). *Multimodal deep learning for early childhood emotion recognition using CNN and LSTM*. Pattern Recognition Letters, 146, 150–157. https://doi.org/10.1016/j.patrec.2021.03.010.