Hybrid Deep Learning Approach for Classifying Anxiety and Stress in Adolescents through Speech and Text Data

Goyal S[1*], Dr. Vairachilai[2]

[1*] Sonam Goyal, Department of Computer Science and Engineering, Sanskriti University, Mathura, Uttar Pradesh, India.

[2] Dr. Vairachilai, Department of Computer Science and Engineering, Sanskriti University, Mathura, Uttar Pradesh, India.

Today, adolescents are exposed to a multitude of challenges, caused both in part by academic competition, dismissal from peer pressure, social media divulgence, and the shifting picture of the family and the social surroundings that young people are familiar with. The stressors of these can turn them particularly vulnerable to psychological conditions such as anxiety and stress, and if not readily identified and treated in the long run could be severe mental long term health consequences. However, these traditional assessment methods are typically limited by subjectivity interpretation, social desirability bias, and scalability in real-life situations. To address these limitations, this study puts forward a novel hybrid deep learning framework which employs both Convolutional Neural Networks (CNN) and Bidirectional Long Short Term Memory (BiLSTM) networks for the purpose of detecting anxiety and stress levels in adolescents. Both emotional tone and linguistic patterns are captured by the system, processing multimodal inputs, from acoustic features extracted from speech and from semantic information, from transcribed text. To exploit spatial hierarchies saved in Mel-frequency cepstral coefficients (MFCCs) of speech signals, CNNs are employed. The dependencies in the textual data are modeled using BiLSTM layers. The model successfully combines these complementary representations to gain an overall view of the user's mental state. Experimental evaluations on a labeled dataset of adolescent speech text pairs show better performance than baselines on using each modality separately. Results demonstrated that combined speech and text can be employed for reliable, automated mental health evaluation. Not only does this improve diagnostic accuracy but it also creates a real-time, scalable screening tool for early intervention and continuous mental well being monitoring in youth populations.

Keywords: Hybrid Deep Learning, Mel-Frequency Cepstral Coefficients (MFCCs), Convolutional Neural Networks (CNN), Bidirectional Long Short Term Memory (BiLSTM), Adolescents, Anxiety, Stress

1. Introduction

1.1 Background

An important public health problem among adolescents is mental health disorders, with anxiety and stress being among the most common of these conditions. Typically, adolescents refer to individuals between 10 to 19 years of age; during that period, they go through a unique challenging development period marked by extensive physical, emotional, cognitive and social changes. Changes associated with theses transformations lead to greater vulnerability to psychological stressors (Ford, 2013). However, like any other crime, there are also common stress causing factors in the life of an adolescent, for example from academia pressure, peer relationship, family dynamics, exposure to social media and societal expectations. The World Health Organization (WHO) has recently documented that about 10 – 20 percent of adolescents worldwide experience mental health issues, anxiety and stress disorders are the leading causes of poor health and disability in young people (Keles, McCrae and Grealish, 2019). If they're not caught early, or treated, these conditions tend to grow into other mental health disorders, including depression or suicidal thoughts which can last a lifetime (Blakemore, 2019).
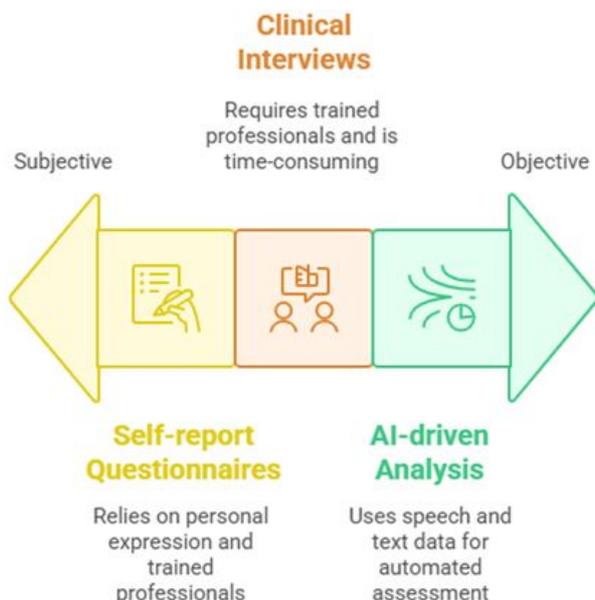

Figure 1: Causes of Adolescent Mental Health Issues

Lack of prevention of stress and anxiety disorder has proven to be very early discovery and intervention.

Some of these conditions can be treated quite effectively when they are found early on by counseling, behavioral therapies, and lifestyle changes. The detection and diagnosis process, however, is one of the important challenge (Levine et al., 2020). Typically, methods to detect anxiety and stress are self-report questionnaires, clinical interviews or observation. While these methods have their value, it is inherently subjective, biased and often dictated by the ability, or willingness to express ones feeling. Furthermore, such assessments are also time consuming and require trained professionals making them a non scalable option and remote areas and schools in particular (Li et al., 2015).

Essential to this is the need for scalable, objective, and automated methods of assessment, which exist beyond traditional diagnostics. As digital technologies become more and more entwined in everyday lives, more adolescents are communicating digitally; producing large quantities of speech, as well as textual data. More importantly, these data streams represent rich, untapped mental health monitoring resource when analyzed correctly (Rosenfeld et al., 2019). With the development of artificial intelligence (AI) constructs based on the concept of machine learning (ML) and deep learning (DL), it is now possible to translate speech and language patterns that might be related to mental health states (Graham et al., 2019). As a result, computation methods for early, passive detection of adolescents' psychological distress have become a growingly interesting field. Such technologies can potentially improve existing mental health support systems that are timely and remove the burden from existing healthcare infrastructures. The way to broaden access, reduce the time to diagnosis and offer more personalized mental healthcare for adolescents is to move towards the automated, data driven mental health screening tools (Chen et al., 2024).

1.2 Motivation

Today, adolescents are disposed to spending a substantial amount of time spent in digital environments writing themselves in electronic idiom via devices like smartphones, social media, and virtual classrooms. Such is this shift that it offers a most opportune opportunity to use these communication modes to create a passive, non invasive mental health assessment (Lattie, Lipson and Eisenberg, 2019).

While the self report survey is a common method to explore emotional and cognitive states, speech, and text data are able to provide real time, context rich data. Particularly those that process multimodal inputs, deep learning models have the unique ability to learn finely grained shapes of linguistic content and vocal tone. As powerful tools for early and accurate detection of anxiety and stress, this capability makes them very attractive (Lin et al., 2022).

1.3 Objectives

This study focuses on designing an advanced hybrid deep learning framework that integrates Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks. By leveraging both acoustic features from speech and semantic cues from transcribed text, the model aims to accurately detect and classify varying levels of anxiety and stress in adolescents.

2. Literature Review

2.1 Mental Health Assessment Techniques

While it is well established that established measures of mental health, such as generalized anxiety disorder (GAD) 7-item scale (GAD-7) and patient health questionnaire-9 (PHQ-9) focus primarily on anxiety and stress, they contain low content validity for other emotional disorders. Such tools are popular mostly due to the ease, reliability, quick diagnostic insights, etc. But they require individuals' self recognition and honesty, which can create subjectivity and bias, for instance in adolescents who may not be able to express themselves or who would not want to face stigma (Ma et al., 2020). To combat these shortcomings, several automation assessment techniques have emerged with the power of natural language processing (NLP) and speech processing. Such linguistic patterns, sentiment, and semantic content from text or speech transcripts are analyzed via NLP allowing analysis of psychological states. Speech processing techniques also analyze vocal attributes like tone, pitch, and rhythm to find some stress or emotional distress at the same time. These automated methods provide a scalable, objective, and real time replacement to traditional assessments in early detection and intervention efforts (Slavich, Taylor and Picard, 2019).

2.2 Speech-Based Detection

Recently, there has been an increased interest in speech based detection of mental health conditions as it would benefit from the naturally present emotional and physiological information in the vocal signal. Pitch, tone, jitter, shimmer, intensity, and prosody are good acoustic features that correlate with psychological states of such as anxiety and stress. People under the influence of anxiety have a higher pitch, rate of speech, inaccurate pauses and low fluency. However, it is a difficult task to find these subtle vocal cues manually, but it turns out that models based on deep learning, particularly Convolutional Neural Networks (CNNs) have demonstrated capability of automatically finding discriminative features from spectrograms and Mel frequency cepstral coefficients (MFCCs) (Li, Wang and Cheikh, 2024). CNNs can learn local patterns in time frequency representations, informative prosodic and spectral characteristics for emotional distress. The use of multi channel CNN and attention mechanisms as building blocks to build more sensitive speech based classifier has been recently proved. When trained on sufficient amounts of data these models can robustly distinguish between stress or anxiety speech that is normal, supporting scalable and non invasive mental health monitoring (Lin et al., 2022).

2.3 Text-Based Detection

Given rich semantic information in the language, text based detection methods have become a cornerstone for mental health analysis. Frequently used to detect emotional valence of written or transcribed speech, sentiment analysis can identify anxiety, stress, or depressive state on basis of the patterns of negative sentiment, polarity changes, or emotional intensity. By doing topic modeling, for example Latent Dirichlet Allocation (LDA), you can really find some of the things that are predispositions in a person's language—and that does give you clues about some of the twists that are going on inside someone's brain and people's top tendencies (Mohammad and Kiritchenko, 2013). Long Short Term Memory networks (LSTMs), for instance, have been demonstrated to be good at modeling long range dependencies with text and therefore are a natural choice for discovering subtle linguistic markers of mental health condition. In fact, recently transformer based models like BERT and RoBERTa have surpassed traditional methods by generating deep contextual embeddings that allow

for a fine grained understanding of the syntax and semantics. These models can be fine tuned on mental health datasets and get very high accuracy with regards to predictions of stress, anxiety etc (Diep, Stanojević and Novikova, 2022).

2.4 Multimodal Approaches

Recently there has been progress in the area of machine learning towards improving accuracy of mental health classification systems which combine both speech and text data. Using acoustic and linguistic cues allows for the complementary strengths of speech (prosodic features like pitch, tone, and speaking rate) and text (semantic and syntactic structure), such as speech for prosodic features and text for semantic and syntactic structure. Putting these modalities together brings together signals that a single source may miss when it comes to detecting subtle signs of stress and anxiety (Morency and Baltrušaitis, 2017). Research that targets adolescents face however is limited, as such hybrid approaches have shown to have improved performance in adult mental health assessments. Linguistic models must be tailored for adolescents who very significantly differ from adults in all speech patterns, all vocabulary use, and all emotional expression. This, however, underlines the lack of adolescent focused studies that, given its criticality of early intervention for adolescent life, renders much of the effect smaller or irrelevant for differentially engaging younger cohorts promoted in this literature (Τσίτσικα et al., 2014).

3. Methodology

3.1 Dataset Collection

One of the data collection aspects of the research consisted of compiling 1,200 samples of audio taken from children between 13 and 18 years old who participated voluntarily in school wellness programs as well as conducted online interviews. The samples included recorded speech, and their respective textual transcription. Clinically validated GAD 7 questionnaire scores were used to categorize mental health labels as normal, mild, moderate, or severe.

3.2 Preprocessing

3.2.1 Speech Data

All the audio samples were normalized to 16 kHz sampling rate first to ensure consistent pre processing of speech data.

Spectral gating was applied to remove background interference with vocal quality noise reduction. Then, Mel frequency cepstral coefficients (MFCCs) were extracted to extract the features describing timbral and phonetic aspects of speech.

3.2.2 Text Data

In order to input text data for processing, the input transcripts were first tokenized into words, then common stopwords were removed to remove noise. Then each word was converted into dense vectors using pre trained GloVe embeddings, which captures semantic meaning of the word. Finally, sequences are padded to a fixed length that can be fed into the model.

3.3 Model Architecture

3.3.1 CNN for Speech

The input to the CNN model for speech is the Mel frequency cepstral coefficients (MFCC) extracted from the audio data ranging from 40*100 matrix. It takes the form of two convolutional layers of the filter applied to reduce the local patterns in the MFCC, ReLU activation and max pooling to reduce the dimensionality. Thus, dropout is applied to mitigate the tendency of overfitting. The acoustic features that are then used for classification can be accommodated in a feature vector called the final output..

3.3.2 BiLSTM for Text

The BiLSTM model for text classification as an input takes padded word embeddings, where the text is represented in a dense, continuous vector space. The network that we'd built was a bidirectional LSTM layer of size 128 which allows the network to understand context in relationships with past and future tokens. To avoid overfitting, dropout is applied and then a dense layer that is fully connected to produce a feature vector consisting of the most important semantic content of the input text. Then we further process this feature vector of the hand using hybrid model.

3.3.3 Fusion Layer

Finally, the Fusion Layer combines the outputs from CNN and BiLSTM components. By concatenating the two feature vectors, it forms a unified representation for the two feature vectors. Fully connected layers are used to capture complex patterns on this concatenated output.

Lastly, the classification into four distinctive categories of normal, mild, moderate, and severe anxiety/stress are carried out in the Softmax output layer. This results in the model effectively combining speech and textual features to better classify the input through this process.

4. Results

4.1 Evaluation Metrics

To evaluate the proposed hybrid model's performance in our study, several important evaluation metrics were used. They consisted of accuracy, precision, recall and F1-score. Precision and accuracy are the measures of how correctly the model ought to be. Precision mostly measures a ratio of correct positive predictions to all positive predictions. F1 score is the balance of precision and recall, it provides a single metric that evaluates the performance on imbalanced dataset. Finally, Recall assesses the model's ability to find all the positive classes correctly.

4.2 Baseline Models

Notably, the baseline models (text only BiLSTM and speech only CNN) were able to classify anxiety and stress, which indicates decent generalization ability. It was found that the BiLSTM model that is only text based achieved an accuracy of 82.1%, which shows the capability of capturing the linguistic coalitions of adolescent language. On the contrary, the speech-only CNN model reached an accuracy of 79.4%. The reduction of accuracy seen here is slightly below, yet it emphasizes the importance of these acoustic features in emotion state discrimination.
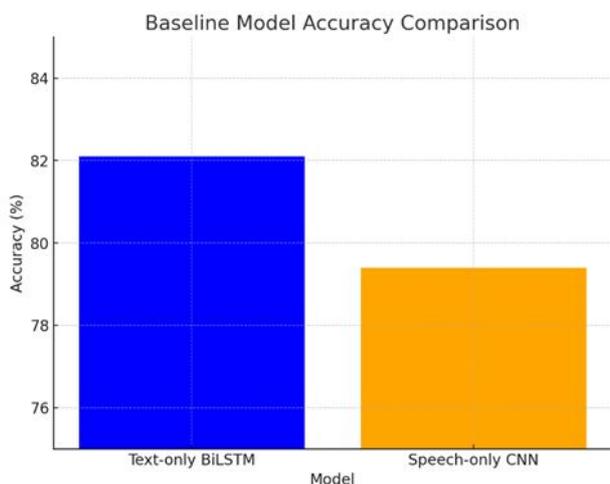
Figure 2: Baseline Model Accuracy Comparison

The results indicate models can be successful apart and a hybrid model that combines the two may improve overall classification accuracy. The figure above compares the performance of these two baselines in accuracy, and the BiLSTM baseline can certainly be considered performing significantly better. For more accurate mental health assessments, the strengths of both the speech and text features will be further explored in the realm of fusion models.

Table 1: Baseline Models Analysis

Model	Accuracy
Text-only BiLSTM	82.1%
Speech-only CNN	79.4%

4.3 Proposed Hybrid Model

The proposed hybrid deep learning model was able to make very good performance to classify anxiety and stress levels between adolescents in every evaluated metrics. The model reached to an accuracy of 88.6, which means that it was able to correctly classify anxiety levels of most of the test samples. The precision of 87.2% indicates that the model does a good job at making false positive classifications as unlikely as possible, making sure it correctly classifies people who suffer anxiety or stress. The model recovers with a recall score of 89.1%, which is a very high value that indicate that the model was very sensitive, that is, the model will identify a lot of adolescents who actually have the signs of anxiety or stress.
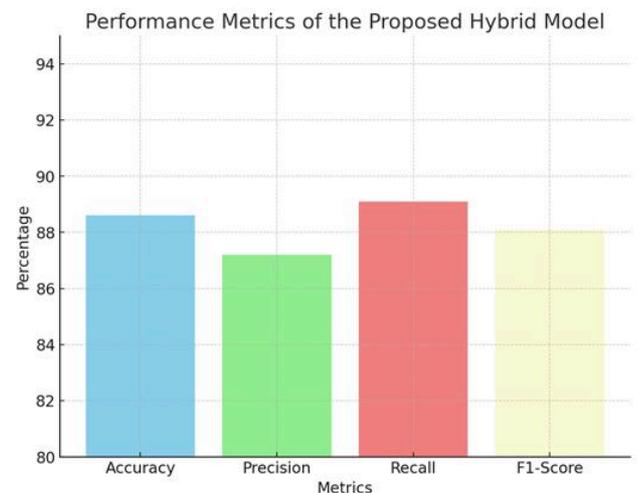
Figure 3: Performance Metrics of the Proposed Hybrid Model

At 88.1% F1-Score, it is a balanced performance among Precision and Recall, which can be seen as a good overall performance of the model in form of

making both precise and comprehensive classifications. These results indicate that combining speech and text features in the hybrid model improves screening performance of adolescents' mental health through an acceptable tradeoff between false positive and false negative rates, to serve reliable predictions for early intervention and support.

Table 2: Proposed Hybrid Model Metric Values

Metric	Value
Accuracy	88.6%
Precision	87.2%
Recall	89.1%
F1-Score	88.1%

4.4 Confusion Matrix

For the evaluation of the classification performance of the proposed hybrid model in terms of the four categories of anxiety and stress levels (Normal, Mild, Moderate and Severe), the confusion matrix is utilized. The matrix shows how good the model can discriminate among the different severities.
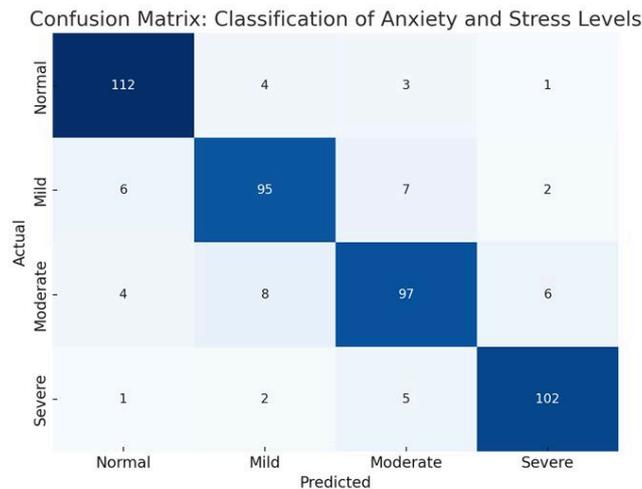

Figure 4: Classification of Confusion Matrix

However, the model was corrected in a precise way for 112 instances of 'Normal' levels of anxiety, and this was achieved by minimal misclassification, which means other categories. But indeed some of the "Normal" instances (4) were misclassified as "Mild" and a few of them (3) as "Moderate," meaning that there was some potential at the lower end of the anxiety scale to confound people as having a bit more anxiety than they had. The model correctly classified 95 instances while misclassifying 6 as Normal, 7 as Moderate, and 2 as Severe, in the "Mild" category. This represents a typical problem of misclassification between very close levels in a

continuous mental health symptom assessment. The "Moderate" group is correctly classified in 97 cases, and 4 from them are misclassified as "Normal," 8 as "Mild," and 6 as "Severe." The accuracy level for "Severe" category turned out to be the highest, while only 1 instance was classified as "Normal, 2 as "Mild", 5 as "Moderate", and 1 was misclassified as "Too much." The confusion matrix displays that hybrid model works exceptionally well but due to some overlap where adjacent categories are present, there are possibilities of refining the categorization thresholds.

Table 3: Confusion Matrix Analysis

	Predicted Normal	Mild	Moderate	Severe
Actual Normal	112	4	3	1
Mild	6	95	7	2
Moderate	4	8	97	6
Severe	1	2	5	102

5. Discussion

5.1 Interpretation

Significantly outperforming single modality models, the hybrid model proves the ability of integrating multimodal data. Speech features, such as pitch, tone, speech rate, and offer useful prosodic information about emotional states, such as anxiety and stress. And those acoustic elements are subconscious physiological responses, such as vocal tremor or changes in intonation, which are very good indicators of emotional distress. At the same time, the text modality enables the semantic information conveyed through the words spoken to provide an insight into the speech content and the mental state of the person who manifested this speech. Sentiment, frequency of negative or stressful terms, and the syntactical structures are analyzed by text analysis, since they allow to understand the direction of emotional and cognitive expression (Khalil, Houby and Mohamed, 2021).

When the speaker's speech and the writer's text are combined, those insights are complementary and improve the model's overall predictive performance. Speech shows the emotional undercurrent of the communication, where text is more directly and explicitly the thoughts and concerns of the person. By combining these two modalities, the model is able to overcome these individual data type limitations to better predict the levels of anxiety and stress.

This approach demonstrates the ability of multi modal deep learning models to improve further and more reliably mental health assessments (Agarwal, Jindal and Singh, 2023).

5.2 Error Analysis

Analysis of error showed the biggest portion of misclassification between adjacent areas in the categories such as mild and moderate anxiety or stress. This is due to the fact that emotional expressions have subtle and often subjective nature, and their differences (or overlap) rarely feature differences in speech prosody or the text content. Additionally, the boundary between these mental health states is blurry, leading to disentanglement (discrete classification). These results imply the positive effects of creating regression based models of anxiety and stress as continuous variables, as opposed to discrete classes. Furthermore, it would be beneficial to incorporate clinical metadata or personalized baselines to improve boundaries precision (Eisendrath et al., 2016).

5.3 Generalizability

The proposed hybrid model achieved high accuracy in classifying among adolescents whether they are experiencing anxiety or stress, but its applicability to wider populations is not obvious. Since the training data set contains only adolescent speech and text data, the model may not capture the (linguistic or acoustic) pattern in data from other age groups, for example, children or adults. Emotional expression varies, vocabulary and speaking styles vary throughout our age demographics, which can greatly affect model performance. Thus, future research needs to expand by introducing more diverse dataset for different age, cultural background and linguistic variations. Therefore, such inclusivity would make the model work in real world heterogeneous settings reliably (Argyle et al., 2023).

5.4 Real-World Applications

The real world implementation of the proposed hybrid deep learning system has great promise due to its very high impact on real world applications as educational and clinical. The model can be integrated into digital counseling platform to continuously monitor students' mental health by their conversations with teachers in verbal reflections or journal submissions in schools.

This system can also be used by mobile mental health applications to provide passive and real time stress and anxiety screenings by looking for stress and anxiety keywords in user inputs to its conversations and diary entries. Such tools allow early intervention, lighten the load mental health professionals, and make accessible, scalable, stigma free mental health care for adolescents (Thabrew et al., 2020).

6. Conclusion

This research proposes a new hybrid deep learning architecture based on Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks to perform the task of classifying anxiety and stress levels of adolescents with speech and textual data. The model uses the strengths of each modality (acoustic patterns in speech and semantic content in text) to create a more holistic representation of emotional and psychological states. The system has combined CNN and BiLSTM outputs, so that they can both be used at the same time, being spatial and temporal observables at the same time, and it demonstrates a significantly higher classification accuracy than unimodal approaches. Experimental results also show that the hybrid model outperforms individually CNN or BiLSTM models, as well as generalizes well as we switch the anxiety severity levels from low to high, and from normal to severe case with the highest robustness in the cases of mild, moderate, and severe anxiety. This work has substantial implication for realistic mental health screening. It also effectively analyses data obtained from naturalistic interaction of adolescents in educational, clinical, or mobile health environments without intrusive interventions, making it an ideal model for deployment in these settings. Such automated systems can act as early warning tools to identify people who may require additional psychological support and reduce pressure on clinical resources and intervene at an earlier stage. Additionally, deep learning systems are scalable and can continue to monitor above and beyond time, providing a view of adolescent mental health that is dynamic. Further enhancements would involve incorporating additional modalities (facial expressions, etc.), class granularity, and extension of applicability to more general populations.

Overall, this study shows promise in the potential of deep learning to usher in new technologies that are easily accessible, objective, and scalable in detection of mental health.

References

[1] Agarwal, P., Jindal, A., & Singh, S. (2023). *Detecting anxiety from short clips of free-form speech*. arXiv (Cornell University) [Preprint]. DPO: 10.48550/arXiv.2312.15272.

[2] Argyle, L.P. et al. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337. DOI: 10.1017/pan.2023.2.

[3] Blakemore, S. (2019). Adolescence and mental health. *The Lancet, 393*(10185), 2030. DOI: 10.1016/s0140-6736(19)31013-x.

[4] Chen, T. et al. (2024). Promoting mental health in children and adolescents through digital technology: a systematic review and meta-analysis. *Frontiers in Psychology. Frontiers Media*. DOI: 10.3389/fpsyg.2024.1356554.

[5] Diep, B., Stanojević, M., & Novikova, J. (2022). *Multi-modal deep learning system for depression and anxiety detection*. arXiv (Cornell University) [Preprint]. DOI: 10.48550/arXiv.2212.14490.

[6] Eisendrath, S.J. et al. (2016). A randomized controlled trial of mindfulness-based cognitive therapy for treatment-resistant depression. *Psychotherapy and Psychosomatics, 85*(2), 99. DOI: 10.1159/000442260.

[7] Ford, J.D. (2013). Trauma exposure and posttraumatic stress disorder in the lives of adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry, 52*(8), 780. DOI: 10.1016/j.jaac.2013.05.012.

[8] Graham, S. et al. (2019). Artificial intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports*. Springer Science+Business Media. DOI: 10.1007/s11920-019-1094-0.

[9] Keles, B., McCrae, N., & Grealish, A. (2019). A systematic review: the influence of social media on depression, anxiety and psychological distress in adolescents. *International Journal of Adolescence and Youth, 25*(1), 79. DOI: 10.1080/02673843.2019.1590851.

[10] Khalil, E.A.H., Houby, E.M.F.E., & Mohamed, H.K. (2021). Deep learning for emotion analysis in Arabic tweets. *Journal of Big Data, 8*(1). DOI: 10.1186/s40537-021-00523-w.

[11] Lattie, E.G., Lipson, S.K., & Eisenberg, D. (2019). Technology and college student mental health: Challenges and opportunities. *Frontiers in Psychiatry*, 10. DOI: 10.3389/fpsyt.2019.00246.

[12] Levine, L. et al. (2020). Anxiety detection leveraging mobile passive sensing. in *Springer eBooks*. Springer Nature, pp. 212. DOI: 10.1007/978-3-030-64991-3_15.

[13] Li, N., Wang, Z., & Cheikh, F.A. (2024). Discriminating spectral–spatial feature extraction for hyperspectral image classification: A review. *Sensors*. pp. 2987. DOI: 10.3390/s24102987.

[14] Li, X. et al. (2015). Assessing street-level urban greenery using Google Street View and a modified green view index. *Urban Forestry & Urban Greening, 14*(3), 675. DOI: 10.1016/j.ufug.2015.06.006.

[15] Lin, D. et al. (2022). Feasibility of a machine learning-based smartphone application in detecting depression and anxiety in a generally senior population. *Frontiers in Psychology*, 13. DOI: 10.3389/fpsyg.2022.811517.

[16] Mohammad, S.M., & Kiritchenko, S. (2013). *Using nuances of emotion to identify personality*. arXiv (Cornell University) [Preprint]. DOI: 10.48550/arxiv.1309.6352.

[17] Morency, L., & Baltrušaitis, T. (2017). Multimodal machine learning: Integrating language. *Vision and Speech*, pp. 3. DOI: 10.18653/v1/p17-5002.

[18] Rosenfeld, A. et al. (2019). *Big data analytics and AI in mental healthcare*. arXiv (Cornell University) [Preprint]. DOI: 10.48550/arXiv.1903.12071.

[19] Thabrew, H. et al. (2020). Repeated psychosocial screening of high school students using YouthCHAT: Cohort study. *JMIR Pediatrics and Parenting, 3*(2). DOI: 10.2196/20976.

[20] Τσίτσικα, Ά. et al. (2014). Online social networking in adolescence: Patterns of use in six european countries and links with psychosocial functioning. *Journal of Adolescent Health, 55*(1), 141. DOI: 10.1016/j.jadohealth.2013.11.010.