International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

A Critical Narrative Review of Machine Learning-Based Student Mental Health Assessment: Classification, Clustering, and Explainable AI

Khushal Rajani (Assistant Professor)
Department of Computer Science, Smt. Chandramaniben Zaverchand Meghji Gosrani BCA College, Oshwal Education Trust, Jamnagar, Gujarat, India
ORCID: https://orcid.org/0009-0009-9392-2461

***Abstract*- student mental health is a public health and educational concern, as depression, anxiety, and chronic stress lead to poor academic performance, dropout, and risk for self-harm. Traditional approaches to assessment including self-report questionnaires and clinical interviews face important barriers like cost stigma bias and generalizability. Novel applications of machine learning (ML) show great potential to enable scalable screening, risk prediction and student segmentation. Methods this narrative review covers ML-based approaches to student mental health assessment through supervised classification, unsupervised clustering and model-agnostic explainable AI (XAI), in 22 articles from 2013-2026 (17 classification and prediction studies, four clustering studies and one association analysis). Results ensemble algorithms, in particular Random Forest and other decision-tree classifiers, show promising predictive ability. Clustering algorithms support hypothesis-free student profiling. XAI methods, such as SHAP and LIME, support model transparency. Limitations include small and unrepresentative datasets, noisy labels, lack of external validation, algorithmic bias, data privacy concerns and poor clinical utility. The review emphasizes that the development of a trustworthy, ethical, and effective ML-based student mental health assessment system must include strong validation, transparency, fairness, and user-centred explainability.**

***Index Terms*—Student mental health, machine learning, classification, clustering, explainable AI, depression, anxiety, stress, SHAP, LIME, educational data mining.**

I. INTRODUCTION

Students in transition to adulthood results in the most common disorders a survey conducted by the World Health Organization in multiple countries of first-year college students that the majority were positive to have at least one common disorder in their lifetime with depression and anxiety the most common. [1] Untreated disorders but affect their class attendance performance social lives; but stigma and lack of resources often deter students from seeking out a counselor.

Conventional measurement includes validated questionnaires like the Patient Health Questionnaire-9 (PHQ-9) [2], the Generalized Anxiety Disorder-7 (GAD-7) [3], and the Perceived Stress Scale (PSS) [4]. These are administered episodically and depends on the participant's honesty. They are valid only for a specific point in time and do not track longitudinal trends. ML has been used for mental health broadly [5][7], as well as more in particular, in students [8][10]. In these prior studies, models have used questionnaire responses, previous test results, smartphone sensing, and language [9]. Techniques include supervised classificationwhere models are trained on data with known labelsas well as unsupervised clusteringwhere unseen types of subjects are discovered. In the field, trained models should be interpretable by counselors and students alike [11], [12].

While most recent reviews about students classification of depression, anxiety, and stress [13][15], out of our (non exhaustive) collection only one consider clustering and XAI (explanaibility); see Section II-C. We present (1) a taxonomy of data sources and tasks, (2) a synthesis of classification and clustering (including evaluation protocols, and a count of reported practices in the works we consider), (3) an analysis of XAI application and misapplication, (4) open challenges and a research agenda. Section II outlines the way, and Sections IIIVII present the background, data, classification, clustering, and XAI; Sections VIII and IX discuss challenges and directions, and X presents final remarks.

II. REVIEW METHODOLOGY

The paper is a narrative review, structured as a template aligned to the transparency guidance of PRISMA 2020 [16], but it is not a systematic review (focused searches, single screening step, no meta-analysis, no formal risk-of-bias assessment). This review is configured around 4 high-level research questions (RQ): RQ1 on what data sources and features are being used, RQ2 on what classification/clustering algorithms are used and how they are being evaluated, RQ3 on how XAI is being implemented and how the explanation quality is being assessed, RQ4 on limitations and ethical issues that are limiting deployment.

A. Search and Selection

Sources searched (November 2026) were publisher and repository platform search A-Z indexed by web-search (PubMed Central ScienceDirect SpringerLink MDPI the ACM Digital Library, proceedings indexed in IEEE, preprint servers) for studies published 20132026. Four searches were conducted for classification, clustering, explainability, and passive sensing; a representative search (in title/abstract) is ('student' OR 'university' OR 'college') AND ('mental health' OR 'depression' OR 'anxiety' OR 'stress') AND ('machine learning' OR 'clustering' OR 'explainable'). An additional search was run using four topic search strings against the Consensus academic search engine (which indexes Semantic Scholar, OpenAlex, and PubMed, but preprints are de-indexed) and setting to exclude preprints. Title/abstract screening was completed by one reviewer, and the full text of all included searches was not screened; the results available from this supplementary source are reported as described in the abstracts. The flow of study selection is summarized in Table I.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

TABLE I. STUDY SELECTION FLOW

Stage	Records	Notes
Supplementary search (four queries returning 30, 20, 15, and 12 records, with overlap)	77	Preprints excluded by the search filter
Retained after single-reviewer title and abstract screening	33	Full texts not reviewed
Supplementary sources in the synthesis	20	16 primary studies (Table IV); four reviews or methodological overviews
Core evidence base from the main search	9	Three systematic reviews and six primary studies (Table III); record counts for the main search were not logged
Total sources in the synthesis	29	22 primary studies (GLOBEM counted as one) and 7 reviews or methodological overviews

B. Criteria and Extraction

We included articles if they (i) studied assessment, prediction, or clustering of student behaviour in schools, colleges, or university, (ii) incorporated ML, or demonstrated how information from behavioral sensing data could benefit ML for assessment, prediction, or clustering, and (iii) were English peer-reviewed journal or proceedings articles. We did not include preprints except for a small number of preprints on arXiv that provide useful methodological context. Most of the evidence base is comprised of 6 primary research articles and 3 systematic reviews (Table III). We include general-population social-media articles [9], [10] solely to motivate text-based methods, but these are not part of the evidence base. Articles that lacked detail on methods used, or that were secondary papers based on clinical populations, were excluded and editorials that did not contain any empirical data.

No formal risk-of-bias assessment was performed as most abstracts lacked enough detail for risk-of-bias assessment with a tool such as PROBAST [17]. Reporting practices related to risk of bias (validation design metrics calibration, significance testing, XAI application) were tabulated in Table V. A risk of bias assessment using PROBAST for each of the referenced studies in this review can be an avenue for future research.

C. Limitations of This Review

Search limitations not complete yet because was limited to online availability and might not have captured everything used single screen in extra search used only 1 engine (not ieee Xplore Scopus, PubMed) for extra search only screened abstracts (not pre-registered) not screening non-English language papers and some of the papers included are conference papers or have less robust peer review. Known reported accuracies obtained from abstracts/tables and not verified that means, counts are approximate only; comparisons are this way purely descriptive. Formal systematic review would include double screen, double inclusion, full-text inclusion and risk of bias assessment.

III. BACKGROUND

The most frequently modeled, measured conditions are depression, anxiety, and stress; and are then followed by burnout loneliness sleep disturbance, suicidal thoughts and psychological distress. Most labels are based on questionnaire cut-offs (e.g. PHQ-910 [2], GAD-7 10 [3]) and not clinical diagnoses; so most models predict a positive screening result rather than a diagnosis of the disorder. Learning ways: the standard-supervised (features to labels), the unlabeled data (structure in unlabeled data), semi-supervised and self-supervised; explaining models (XAI) such as post hoc explanation tools, intrinsically explainable or interpretable models or analysis; and examples of the eight steps in the machine learning pipeline: data sources (surveys, sensors), data preprocessing (data cleaning, data encoding, train/test split), feature selection (filter methods, embedded methods), modeling (classification, clustering) and tuning, model validation (preferably external and/or temporal validation), model explanation (e.g. SHAP LIME), counselors (human in the loop), and drift/fairness monitoring. The pipeline is depicted in Fig. 1.

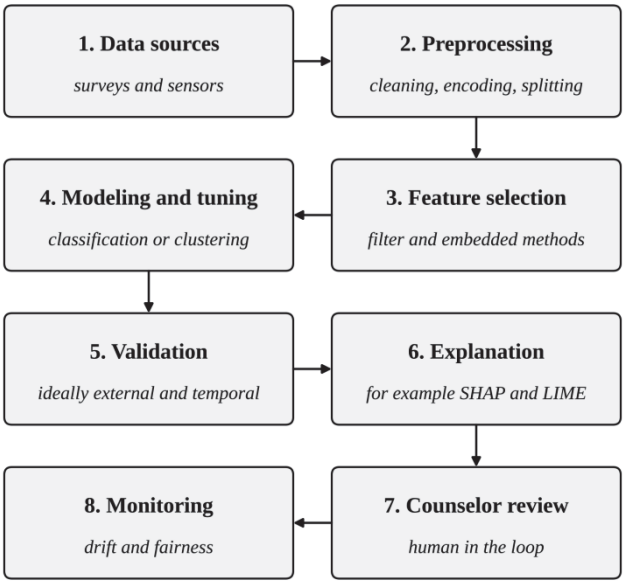


Fig. 1. Typical machine learning pipeline for student mental health assessment.

IV. DATA SOURCES, FEATURES, AND PREPROCESSING

A. Survey and Academic Data

Most common source is cross-sectional survey data with validated scales, like PHQ-9, GAD-7, PSS, and DASS-21, and demographic lifestyle academic, and socio-economic data. These are cheap, small, and self-reported, and mostly from one university. Public datasets can be reused for many papers, but when one dataset underpins quite a few papers, the papers cannot be deemed independent evidences. Attendance, grades, and learning-management-system logs are passive, scalable behavioral proxies of mental state, but only indirectly and also there are privacy issues.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

B. Passive Sensing

StudentLife already demonstrated that smartphone-sensed activity conversation sleep, location are related to depression, stress, and performance [8], and GPS and phone usage features quantify depression symptom severity [18]. A multi-year passive-sensing benchmark for depression detection (GLOBEM) found little generalization of proposed algorithms [19], [20]. A symptom-specific model of paper-based depression symptom severity using data from phones and wearable devices achieved 81.5% recall and 69.1% precision on weekly depression status in 83 undergrads [21], and 85.7% accuracy predicting end-of-semester symptoms in 138 students sometimes 11-15 weeks before the end of the semester [22], with improvements from contextually filtered, more personalized features over population means [23], [24]. We find small samples, short monitoring periods, and limited external validation for a recent broad review of passive sensing (clinical and population samples, non-student) [25], [26].

C. Text, Speech, and Social Media

Language shows affect: initial work measured mental-health indicators in social media [9], [10], and can be reconfigured to transformer models (e.g. [27]) applied to posts, journals, or even therapy transcripts. But the initial detection work wasn't on students; we found no paper doing transformer-based detection for students at all. Also for voice and face, while less common, we found no machine learning papers for students in that data domain either.

D. Preprocessing and Feature Engineering

We found that most methods fall into a set of common strategies: imputation encoding normalization, or dimensionality reduction. For class imbalance, which is often present due to the relatively small number of severe cases, oversampling (e.g. SMOTE [28]), class weighting, or threshold tuning was implemented; for feature selection, which was often performed given the large number of features filter, wrapper, or embedded methods were used. A mistake often made in biomedical ML is that oversampling or scaling is performed before data partition into training and test sets, causing a leak of information and overestimating performance [29]. Table II summarizes data processing.

TABLE II. DATA MODALITIES USED IN STUDENT MENTAL HEALTH ASSESSMENT

Modality	Typical features	Strengths	Limitations
Surveys/ questionnaires	PHQ-9, GAD-7, PSS, demographics, lifestyle	Validated, easy, low cost	Self-report bias, cross-sectional, small samples
Academic/ LMS records	Grades, attendance, login patterns	Passive, scalable	Indirect proxy, privacy issues
Smartphone/ wearable sensing	Mobility, sleep, activity, call/app usage, HRV	Continuous, objective	Missing data, device variance, consent
Text/ social media	Linguistic cues, sentiment, embeddings	Rich affective signal	Platform bias, ground-truth uncertainty
Speech/ video	Prosody, facial cues	Non-verbal signal	Intrusive, high privacy risk

V. SUPERVISED CLASSIFICATION APPROACHES

A. Models

Logistic regression remains a powerful and interpretable baseline, while support vector machines [30] are adapted to small, high dimensional survey data. Random forests [31] was the algorithm most cited across the papers we reviewed (nine out of 12 tabular papers listed in Table V), and was most often cited as the best model. The second most common best model was gradient boosting variants such as XGBoost [32] (most implementations used scikit-learn [33]). Deep models appeared less frequently in the abstracts we examined. LSTMs [34] and transformers like BERT [27] have been shown as appropriate for, and effective on, longitudinal sequences and text respectively, but tend to require more data than typical student studies, and risk overfitting. In the single tabular comparison with multi-layer perceptron [35], simpler models and tree ensembles performed better. The benefit from stacking and similar methods should be tested for significance versus variation in data splits, only addressed by one paper we reviewed [36].

B. Evaluation Practice

Popular metric accuracy not recommended for unbalanced classes. Better metrics are precision recall specificity F1 AUC-ROC, precision recall AUC, calibration; screening normally needs recall but ponder on false positives (alarm, stigma) takes. For validation, stratified, and nested cross-validation, repeated runs and no doubt, temporal, and external validation on another institution needed.

C. Critical Observations

A systematic review of 48 under-graduate papers reported accuracies for stress, 63% to 100%; for anxiety, 53.69% to 97.9%; and for depression, 73.5% to 99.1%; low certainty (GRADE), and the need for external validation [13]. Comparing different data sets, labels, and validation schemes is difficult. Two health sites in Lebanon reported best AUCs, 72.96% to 78.27%, from demographics and general health [35]; a Bangladeshi suicidal-risk paper report best F1-score, 75.36% [37]. Near-ceiling accuracies require diagnosis of label leakage, duplicated records, and features encoding the outcome; many very high accuracy results derive from small, single-site populations with random splitting that artificially boost accuracy estimates of generalizability. Since some questionnaire labels are derived from questionnaires and also as inputs to the predictive model, scale item responses to an input and label should not both be used.

The supplementary search qualifies this, and comparisons must be made within a metric. Of studies that report AUC, the largest sample 61 619 US students [38], ranged from 0.74 to 0.77, and two Lebanese universities ranged from 0.73 to 0.78 [35], but higher values were found for 355 Bangladeshi students (0.87) [39] 1 635 Chinese students (0.87) [40], some 27,901 online records (0.92) [41], and 10,043 Chinese undergraduates (0.93) [42]. Large samples do not automatically yield modest discrimination. The abstracts of the high-AUC cross-sectional studies do not indicate whether their predictors match the label construct, so we flag them as possible leakage candidates, not as successful deployable metrics. Accuracy is 0.570.69 (from 1,697 Bangladeshi students [43]) to 0.79 [40], 0.85 [41], 0.875

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[42], 0.89 [36], 0.897 [39], and 0.90 in 804 students [44]. Although we found that comparison between studies with different sample sizes, label sets, and validation sets is instructive but not conclusive; better comparisons can be found elsewhere: the Lebanese wartime study defined model selection through precisionrecall AUC with Brier scores and calibration curve [45], and another study used repeated stratified cross-validation, Friedman, and Nemenyi tests, and an SMOTE ablation [36]. These are exceptions to the trends summarized in Section V-B.

D. Evidence from Representative Studies

Core evidence base is in Table III, with all additional primary papers shown in Table IV. Three points emerge: high dataset- and feature set-specificity of performance, axial evidence supporting that clustering research is all exploratory, and (3) performance explainability as a later add-on for tabular modelling.

TABLE III. CORE EVIDENCE BASE: REPRESENTATIVE SOURCES IN STUDENT MENTAL HEALTH ASSESSMENT

Ref.	Year	Population and data	Task and methods	Reported findings and caveats
[8]	2014	StudentLife: 48 college students, smartphone sensing over one academic term	Association of sensed behavior with mental health and grades	Showed passive behavioral correlates of mental health; small single-campus sample
[13]	2024	Systematic review of 48 studies on undergraduate students (2,304 records found)	ML detection of depression, anxiety, and stress	Accuracy 63–100% (stress), 53.69–97.9% (anxiety), 73.5–99.1% (depression); certainty of evidence very low (GRADE); external validation needed
[14]	2023	Systematic review, college students	ML prediction of anxiety and stress	SVM and logistic regression reported most effective; age and gender common predictors; most studies did not use variable selection
[15]	2024	Systematic literature review, students, 2011–2024	ML and deep learning for mental health diagnoses	Broad survey of classical and deep models; diagnosis-level conclusions are not specific to student screening
[19], [20]	2022	GLOBEM: four multi-year passive-sensing datasets from university cohorts	Nine prior depression-detection algorithms reimplemented, plus domain-generalization methods	Within-dataset gain over a majority-class baseline only 7.0 ± 7.6%; poor cross-dataset generalization; proposed method gains incremental
[35]	2024	Survey at two Lebanese universities during COVID-19; demographics and self-rated health	Eight models (LR, MLP, SVM, RF, XGBoost, AdaBoost, NB, KNN)	Best AUC: RF 78.27% (depression), NB 76.37% (anxiety), AdaBoost 72.96% (stress); top features self-rated health and age
[46]	2025	4,911 Chinese undergraduates; EPOCH wellbeing survey	K-means, K chosen by silhouette	Two clusters (41.09% and 58.91%); silhouette about 0.3; single university, convenience sample, cross-sectional
[37]	2026	Bangladeshi university students; StudentSafe dataset	Suicidal-risk prediction with several models; SHAP and LIME	XGBoost best, F1 75.36%; key determinants: device usage, institution type, academic achievement, family environment
[47]	2022	Mental health data from 1,000 students at one school	K-means with optimized iteration	Three clusters (20.6%, 31.9%, 47.1%); single-school data

TABLE IV. ADDITIONAL PRIMARY STUDIES FROM THE SUPPLEMENTARY SEARCH (ABSTRACT-LEVEL DATA)

Ref.	Year	Population and data	Task and methods	Reported findings and caveats
[43]	2025	1,697 students at two Bangladeshi public universities; DASS-21	Seven ML models	Best accuracy 0.57 (SVM, depression), 0.69 (LR, anxiety), 0.67 (CatBoost, stress); moderate AUC
[38]	2024	61,619 students, 133 US institutions	XGBoost, RF, DT, LR; 90:10 split	AUC 0.74 and 0.77 for the two outcomes; top predictors financial situation, belonging, disability, age
[48]	2024	1,250 Bangladeshi students; GAD-7, PHQ-9, ISI-7	DT, RF, XGBoost	XGBoost best for all three outcomes; social-media survey, cross-sectional
[44]	2024	804 Bangladeshi students; GHQ-12	16 ML models	Extremely randomized trees best, 90.26% accuracy
[39]	2021	355 students, 28 Bangladeshi universities	DT, RF, SVM, LR; Boruta selection	RF accuracy 0.897, AUC 0.872 for perceived stress
[42]	2025	10,043 Chinese undergraduates; CES-D	Random forest	Accuracy 87.5%, AUC 0.927; cross-sectional
[40]	2025	1,635 Chinese college students	RF, XGBoost, LightGBM, SVM; SHAP	RF best, AUC 0.87, accuracy 0.79
[45]	2026	225 students, one Lebanese university, 2024 war	Nine algorithms; SMOTE vs. balanced pipelines; PR-AUC, Brier, calibration	Balanced RF PR-AUC 0.93 (depression) and 0.88 (anxiety); exploratory, small sample
[36]	2026	Two datasets (777 and 1,100 students); academic stress	Five classifiers; repeated stratified CV; Friedman/Nemenyi tests; SMOTE ablation; SHAP	RF accuracy 0.891; significance testing and ablation reported
[49]	2022	9,738 students, Yingshan County, China	Latent profile analysis, then k-means	Three anxiety profiles; k-means gave similar groupings

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

Ref.	Year	Population and data	Task and methods	Reported findings and caveats
[50]	2026	Healthy Minds Network student survey	Socially fair k-means	Clustering-cost disparities across racial groups reduced by up to 30%
[41]	2026	About 27,901 records, OpenML	LR, RF, XGBoost; SHAP, LIME; fairness evaluation	XGBoost accuracy 0.846, AUC 0.920; fairness checked by gender and financial stress
[21]	2018	83 Dartmouth undergraduates, two 9-week terms	Symptom features from phone and wearable sensing	Weekly depression status predicted with 81.5% recall and 69.1% precision
[22]	2021	138 college students	Features from longitudinal phone and fitness-tracker data	85.7% accuracy for end-of-semester symptoms; above 80% 11 to 15 weeks early
[23]	2019	College students; end-of-semester BDI-II	Association-rule contextually filtered features	Average gain of 9.7% over unimodal baseline; replicated on a second dataset
[24]	2021	First-year college students, two datasets	Collaborative-filtering personalized model	Gains of 5.1% accuracy and 5.5% F1 over a state-of-the-art baseline

E. Reporting Practices Across the Included Studies

Table V presents counts of reporting practices in the 22 primary studies to provide empirical grounding for RQ2 and RQ3. Counts reflect only what is stated in the abstracts and tables, so the totals underestimate the actual number of practices used (just not described). "Tabular" studies are the 12 survey-based classification studies, the other five classification studies are the sensing studies [19][24] (GLOBEM is counted once).

TABLE V. REPORTING PRACTICES ACROSS THE 22 PRIMARY STUDIES (ABSTRACT-LEVEL, APPROXIMATE)

Item	Count	Studies
Classification or prediction	17 of 22	[19]–[24], [35]–[45], [48]
Clustering	4 of 22	[46], [47], [49], [50]
Association analysis only	1 of 22	[8]
Random forest among models compared (tabular)	9 of 12	[35], [36], [38]–[42], [45], [48]
XGBoost among models compared (tabular)	6 of 12	[35], [37], [38], [40], [41], [48]
Logistic regression / SVM named (tabular)	5 / 4 of 12	LR: [35], [38], [39], [41], [43]; SVM: [35], [39], [40], [43]
Best model reported: RF / XGBoost / other / not stated (tabular)	6 / 3 / 2 / 1	RF: [35], [36], [39], [40], [42], [45]; XGBoost: [37], [41], [48]; other: [43], [44]; not stated: [38]
Accuracy reported	9 of 17	[22], [24], [36], [39]–[44]
AUC-ROC reported	6 of 17	[35], [38]–[42]
F1 reported	2 of 17	[24], [37]
Precision and recall reported	1 of 17	[21]
PR-AUC, Brier score, or calibration reported	1 of 17	[45]
Significance test of model differences	1 of 17	[36]
Evaluation on more than one dataset (not necessarily external)	4 of 17	[19], [20], [23], [24], [36]
SHAP used	4 of 17	[36], [37], [40], [41]
LIME used	2 of 17	[37], [41]
Fairness evaluation	1 of 17 classification; 1 of 4 clustering	[41]; [50]
User evaluation of explanations with counselors or students	0 of 22	none
K-means used (clustering)	4 of 4	[46], [47], [49], [50]
Number of clusters chosen by silhouette	1 of 4	[46]
Stability analysis or external validation of clusters	0 of 4	none
Cross-method check of clusters	1 of 4	[49]
Country or region (where stated)		Bangladesh 5: [37], [39], [43], [44], [48]; China 4: [40], [42], [46], [49]; Lebanon 2: [35], [45]; United States at least 3: [8], [21], [38]

Three patterns stand out. First, random forests were the most frequently compared and most frequently reported best model, ahead of XGBoost, so a view that gradient boosting dominates is not supported by this sample. Second, only a small minority of studies reported calibration, significance testing, or evaluation on more than one dataset, which is consistent with the very low certainty of evidence found by the GRADE assessment [13]. Third, XAI use is concentrated in a few recent studies, and no abstract reports evaluating explanations with the people who would act on them.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

VI. UNSUPERVISED CLUSTERING APPROACHES

A. Rationale and Algorithms

Clustering is suitable if labels are often missing, or noisy, or if variation like many stress and coping profiles is the real goal, not routing to a known category; clustering can also be used to reveal subgroups for targeted intervention or resource planning. The most common was k-means [51] (all four clustering papers used it) because it is simple; k-medoids and k-prototypes are robust with outliers or mixed data respectively; the aggregation hierarchical clustering creates a nested tree of risk groups; DBSCAN [52] can find arbitrarily-shaped groups and flag odd students; Gaussian mixture models can give soft groupings and allow symptom severity to be measured on a spectrum. We listed other unsupervised learning algorithms like self-organizing maps, spectral clustering, and deep embedded clustering for very high-dimensional sensing and text data, but these did not appear in the papers we reviewed.

B. Validation and Hybrid Designs

Number of clusters is chosen with silhouette coefficient [53], indices of DaviesBouldin and CalinskiHarabasz, and the elbow method; robustness and external validation with clinical scales or subsequent endpoints offers further support. Cluster centroid interpretation and comparison of clusters are common. Hybrid approaches cluster first and then train cluster-specific classifiers, use membership as a feature, or generate pseudo-labels for semi-supervised learning; although they can boost personalized models, they can also accentuate idiosyncrasies of the clustering process.

C. Critical Observations

Outputs are affected by scaling, features, and initializations; the number of clusters can be selected by trial or clinical acumen; and the clusters' clinical labels are arbitrarily assigned without validation a problematic limitation of all cluster analyses in mental health [54]. Clustering results are a tentative way to describe patterns, not diagnoses, and should be treated as such. A single relevant finding in a sample of 4,911 Chinese undergraduates used K-means with the silhouette criterion (a moderate value at about 0.3) to select two clusters from a convenience sample at one university in a cross-sectional design [46]. The four clustering studies we identified [46], [47], [49], [50] all used K-means profiling; none of their abstracts mention external validation, and they did not externally validate their clusters on clinical outcomes, but cross-method comparisons are informative (latent profile analysis and K-means clustered similar groups in 9,738 students [49]). Although guidelines exist for clustering algorithm selection and evaluation [54], a recent new, fairness-aware K-means algorithm quite a bit reduced disparities in clustering costs across racial groups by as much as 30% [50], an important and unaddressed VIII-C limitation.

VII. EXPLAINABLE ARTIFICIAL INTELLIGENCE

A. Why Explainability Matters

A flag that a student is at risk impacts subsequent counseling, the students' well-being and more. They are welcomed by clinicians, aid them to challenge or validate domain knowledge and are useful to rule out spurious or biased reasoning. They are also related to regulation development: Regulation (EU) 2024/1689 has obligations related to high-risk AI, education uses are part of high-risk AI, inference of emotions is prohibited in educational institutions except when for medical or safety purposes [55], so designers should be able to figure out, what the rules are that apply in their territory. XAI surveys further distinguish between intrinsic interpretability and post-hoc explanation, and between global and local explanations [11], [12], [56]; Rudin suggests that high-stakes decisions be the domain of inherently interpretable models, not post-hoc explanations of black box models [57].

B. Techniques Used

SHAP [58], in which the contribution of each feature to a prediction is derived via Shapley values was most frequently used (4 vs. 2 for LIME, Table V), but by far the most commonly reported type was global importance, dependence, and individual force plots. Local interpretable model explanation (LIME) [59], offers an interpretable model that locally explains the prediction. Model-agnostic techniques include permutation importance and partial dependence/ICE plots while intrinsically interpretable techniques include: regression coefficients, tree rules, and generalized additive models. For deep models, attention, and gradient based saliency techniques like Grad-CAM (gradient-weighted class activation mapping) [60] can provide importance scores for weighted labels, time windows or signal segments. Counterfactual explanations, which quantify the amount by which a feature needs to be changed to achieve a different prediction are appealing as actionable explanations, though this approach has yet to appear in the student papers. Table VI summarizes these techniques.

TABLE VI. XAI TECHNIQUES AND THEIR SUITABILITY

Technique	Scope	Model type	Notes
SHAP	Global and local	Model-agnostic/ tree-specific	Consistent attributions; costly for some models
LIME	Local	Model-agnostic	Instability with correlated features
Permutation importance	Global	Model-agnostic	Biased with correlated features
PDP/ICE	Global/ local	Model-agnostic	Assumes feature independence
Attention/ Grad-CAM	Local	Deep models	Faithfulness debated
Counterfactuals	Local	Model-agnostic	Actionable; feasibility constraints needed
Decision rules/GAM	Global	Intrinsic	High transparency

C. Typical Findings

Top drivers look different according to the dataset: use of device, type of school, academic achievement and family circumstances were the top drivers in a suicide-risk study of Bangladeshi students [37]; self-rated health and age led the way in a Lebanese one [35]. By extending the model interpretation to identify important drivers in this way, we have established what the model is learning about each one cohort, not risk factors for the outcome across populations, and SHAP values are correlation-based, not causal. We have seen several efforts recently to build in fairness into the user evaluation of SHAP

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

and LIME explanations [41], and a systematic review of explainable AI for depression prediction identified only six papers that used SHAP, LIME, or related techniques [61], and none of the student papers we looked at included a user evaluation with counselors or students.

D. Limitations of Current XAI Practice

1) Faithfulness and stability: sampling and correlations among features may change LIME and SHAP explanations and give little insight into the innards of the models.

2) Correlation versus causality: attributions are indicative of model behavior, not causal mechanisms.

3) Lack of evaluation: few studies quantify explanation quality with human- or functionally-grounded evaluations [56], and few involve counselors or students.

4) Audience mismatch: plots used for data scientists may not be actionable for counselors or students.

5) Clusters: explaining clusters is under-explored; typical practice is descriptive profiling rather than surrogate classifiers explained with SHAP.

VIII. CHALLENGES AND ETHICAL CONSIDERATIONS

A. Data Quality, Representativeness, and Label Noise

Most data are small, cross-sectional, single institution convenience samples from specific countries/ fields and because of this some generalisability across countries, gender, and SES tends to be limited. Of the countries we have listed, many primary studies in our review are conducted in Bangladesh (5), China (4), Lebanon (2) and the United States (at least 3) (Table V). The largest samples from the data sets are from the US [38] and China [42], [49]. Given that screening cut-points are by definition imprecant markers of diagnosis, and that there may be under-reporting of certain information because of stigma, the models are training on responses to the questionnaire, not the actual disorder.

B. Overfitting and Weak Validation

Small sample sizes, high feature dimensions, repeated tuning on test data, and no external validation give an overly optimistic estimate of performance. A scoping review of 42 passive-sensing clinical population studies showed 32 (76%) studies with.<100 subjects, 1 (2%) external validation [25], a GRADE assessment of 48 student-studies rated the evidence as very low [13], and nine re-implemented algorithms for depression detection outperformed a single nave baseline by 7.0 7.6% in each dataset and performed poorly on other datasets [19], [20]. Our own calculation was similar: of 17 classification studies, only four had more than one dataset tested, only one reported calibration, and only one tested differences between models for significance (Table V).

C. Fairness, Privacy, and Deployment

Algorithmic bias can bring systematically different error rates for different genders, different races, or socioeconomic groups; and biased proxies caused large disparity in health care algorithm [62]; we found few fairness audits of student algorithms (1 classification [41] and 1 clustering [50]). Passive sensing, academic records, and social media data are sensitive, and re-identification, abuse by institutions, and chilling effects are a concern; federated learning [63], differential privacy, and inference on device are promising techniques but limited use to date. But high accuracy does not necessarily mean good we need to consider how the algorithm is embedded into counseling workflow, escalation path, false positive and false negative handling, human involvement, dataset shift detection. Such systems are screening, not diagnostic, tools that are meant to support not supplant human judgement and not to threaten or bar entry.

IX. FUTURE RESEARCH DIRECTIONS

It aims for: (1) large multi-site longitudinal cohorts with shared metadata for external and temporal validation; (2) interpretable modeling and leakage-free pipelines calibration confidence intervals, statistical model comparison, and formal risk-of-bias assessment using PROBAST [17]; (3) semi- and self-supervised learning on abundant unlabeled data; (4) multimodal fusion with robust missingness handling; (5) personal models capturing one's own baseline; (6) human-centered and responsible XAI with co-designed counterscripts of counselors and students, user studies, and extension to clusters and temporal models; (7) causal inference towards feasible treatment factors; (8) fairness-sensitive and privacy-preserving learning with systematic subgroup tracking; (9) large language models [64] for screening and discussions, with defenses to hallucination, bias, and harms, and with student-specific evidence that we did not encounter in the sources we reviewed; and (10) prospective trials whether ML-assisted screening increases seeking help, results, and cost-effectiveness.

X. CONCLUSION

ML can democratize the early detection of student mental health issues. Our review found, among existing mental health-related clinical ML applications, that tree ensembles (most often, the random forest) were the most common, reported best model for screening depression, anxiety, and stress from questionnaire and behavioral data, and that clustering was effective to find heterogenous subgroups for intervention. Explainable-AI algorithms (e.g. SHAP LIME) are popular but currently under appreciated as the necessary components of trustworthy ML, with studies not at all examining explainability's capacity to inform mental health workers and students in understanding the recommendations generated by ML. The literature suffers from small, biased samples, noisy proxy labels, weak validation, insufficient addressing of fairness, privacy, and clinical utility, and our review was limited by not fully covering the scope of the literature and the processes of summarization and single-reviewer implementation. We are seeking a future of ML-based assessment systems that are realized through collaborations, multi-site, and sufficiently sized data, transparent leakage-free methods, human-centered explanations, and ethically-informed deployment.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

REFERENCES

[1] R. P. Auerbach et al., "WHO World Mental Health Surveys International College Student Project: Prevalence and distribution of mental disorders," J. Abnorm. Psychol., vol. 127, no. 7, pp. 623–638, 2018, doi: 10.1037/abn0000362.

[2] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," J. Gen. Intern. Med., vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[3] R. L. Spitzer, K. Kroenke, J. B. W. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The GAD-7," Arch. Intern. Med., vol. 166, no. 10, pp. 1092–1097, 2006, doi: 10.1001/archinte.166.10.1092.

[4] S. Cohen, T. Kamarck, and R. Mermelstein, "A global measure of perceived stress," J. Health Soc. Behav., vol. 24, no. 4, pp. 385–396, 1983, doi: 10.2307/2136404.

[5] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," Psychol. Med., vol. 49, no. 9, pp. 1426–1448, 2019, doi: 10.1017/S0033291719000151.

[6] S. Graham et al., "Artificial intelligence for mental health and mental illnesses: An overview," Curr. Psychiatry Rep., vol. 21, art. no. 116, 2019, doi: 10.1007/s11920-019-1094-0.

[7] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," Annu. Rev. Clin. Psychol., vol. 14, pp. 91–118, 2018, doi: 10.1146/annurev-clinpsy-032816-045037.

[8] R. Wang et al., "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones," in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp), 2014, pp. 3–14, doi: 10.1145/2632048.2632054.

[9] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. 7th Int. AAAI Conf. Weblogs Social Media (ICWSM), 2013, pp. 128–137, doi: 10.1609/icwsm.v7i1.14432.

[10] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in Proc. Workshop Comput. Linguistics Clin. Psychol., 2014, pp. 51–60, doi: 10.3115/v1/W14-3207.

[11] R. Guidotti et al., "A survey of methods for explaining black box models," ACM Comput. Surv., vol. 51, no. 5, art. no. 93, 2018, doi: 10.1145/3236009.

[12] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Inf. Fusion, vol. 58, pp. 82–115, 2020, doi: 10.1016/j.inffus.2019.12.012.

[13] B. L. Schaab et al., "How do machine learning models perform in the detection of depression, anxiety, and stress among undergraduate students? A systematic review," Cad. Saúde Pública, vol. 40, no. 11, art. no. e00029323, 2024, doi: 10.1590/0102-311XEN029323.

[14] A. Daza, N. Saboya, J. I. Necochea-Chamorro, K. Zavaleta Ramos, and Y. del R. Vásquez Valencia, "Systematic review of machine learning techniques to predict anxiety and stress in college students," Inform. Med. Unlocked, vol. 43, art. no. 101391, 2023, doi: 10.1016/j.imu.2023.101391.

[15] U. Madububambachu, A. Ukpebor, and U. Ihezue, "Machine learning techniques to predict mental health diagnoses: A systematic literature review," Clin. Pract. Epidemiol. Ment. Health, vol. 20, 2024, doi: 10.2174/0117450179315688240607052117.

[16] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," BMJ, vol. 372, art. no. n71, 2021, doi: 10.1136/bmj.n71.

[17] R. F. Wolff et al., "PROBAST: A tool to assess the risk of bias and applicability of prediction model studies," Ann. Intern. Med., vol. 170, no. 1, pp. 51–58, 2019, doi: 10.7326/M18-1376.

[18] S. Saeb et al., "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study," J. Med. Internet Res., vol. 17, no. 7, art. no. e175, 2015, doi: 10.2196/jmir.4273.

[19] X. Xu et al., "GLOBEM: Cross-dataset generalization of longitudinal human behavior modeling," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 6, no. 4, art. no. 190, Dec. 2022, doi: 10.1145/3569485.

[20] X. Xu et al., "GLOBEM dataset: Multi-year datasets for longitudinal human behavior modeling generalization," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS) Datasets Benchmarks Track, vol. 35, 2022, arXiv:2211.02733.

[21] R. Wang et al., "Tracking depression dynamics in college students using mobile phone and wearable sensing," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 2, no. 1, art. no. 43, 2018, doi: 10.1145/3191775.

[22] P. Chikersal et al., "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing," ACM Trans. Comput.-Hum. Interact., vol. 28, no. 1, art. no. 3, 2021, doi: 10.1145/3422821.

[23] X. Xu et al., "Leveraging routine behavior and contextually-filtered features for depression detection among college students," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 3, no. 3, art. no. 116, Sep. 2019, doi: 10.1145/3351274.

[24] X. Xu et al., "Leveraging collaborative-filtering for personalized behavior modeling: A case study of depression detection among college students," Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 5, no. 1, art. no. 42, Mar. 2021, doi: 10.1145/3448107.

[25] S. Shen et al., "Passive sensing for mental health monitoring using machine learning with wearables and smartphones: Scoping review," J. Med. Internet Res., vol. 27, art. no. e77066, 2025, doi: 10.2196/77066.

[26] V. de Angel et al., "Digital health tools for the passive monitoring of depression: A systematic review of methods," npj Digit. Med., vol. 5, art. no. 3, 2022, doi: 10.1038/s41746-021-00548-8.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[28] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002, doi: 10.1613/jair.953.

[29] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," Nat. Commun., vol. 15, art. no. 1829, 2024, doi: 10.1038/s41467-024-46150-w.

[30] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[31] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[32] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[33] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[35] C. El Morr et al., "Predictive machine learning models for assessing Lebanese university students' depression, anxiety, and stress during COVID-19," J. Prim. Care Community Health, vol. 15, art. no. 21501319241235588, 2024, doi: 10.1177/21501319241235588.

[36] A. Omarbekova et al., "An explainable machine learning framework for academic stress classification among university students: A comparative multi-dataset study with ablation analysis and statistical significance testing," Front. Comput. Sci., 2026.

[37] A. Islam, A. Rahman, M. J. Abdin, O. Rahaman, and M. N. Alam, "Addressing the mental health crisis: Understanding suicidal risk factors in university students through interpretable machine learning," in Proc. Int. Conf. Intell. Data Anal. Appl. (IDAA 2025), 2026, doi: 10.2991/978-94-6239-664-7_30.

[38] Y. Zhai et al., "Machine learning predictive models to guide prevention and intervention allocation for anxiety and depressive disorders among college students," J. Couns. Dev., vol. 103, no. 1, pp. 110–125, 2025, doi: 10.1002/jcad.12543.

[39] R. Rois, M. Ray, A. Rahman, and S. K. Roy, "Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms," J. Health Popul. Nutr., vol. 40, art. no. 50, 2021, doi: 10.1186/s41043-021-00276-5.

[40] C. Yu et al., "Machine learning models for predicting the risk of depressive symptoms in Chinese college students," Front. Psychiatry, vol. 16, art. no. 1648585, 2025, doi: 10.3389/fpsyt.2025.1648585.

[41] R. Cernechez et al., "Explainable machine learning for predicting student depression risk," Bioengineering, vol. 13, no. 9, art. no. 1083, 2026, doi: 10.3390/bioengineering13091083.

[42] L. Luo et al., "Predictors of depression among Chinese college students: A machine learning approach," BMC Public Health, vol. 25, 2025, doi: 10.1186/s12889-025-21632-8.

[43] M. E. Hasan et al., "Prevalence, associated factors, and machine learning-based prediction of depression, anxiety, and stress among university students: A cross-sectional study from Bangladesh," J. Health Popul. Nutr., vol. 44, 2025, doi: 10.1186/s41043-025-01095-8.

[44] N. Mumenin et al., "Screening depression among university students utilizing GHQ-12 and machine learning," Heliyon, vol. 10, no. 17, art. no. e37182, 2024, doi: 10.1016/j.heliyon.2024.e37182.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 11 , November - 2026

[45] R. Hoteit, I. Bou-Hamad, and C. El Morr, "Machine learning prediction of depression, anxiety, and stress among university students during wartime in Lebanon," Sci. Rep., 2026.
[46] J. Wan, L. Ren, Y. Tan, Y. H. Wong, C. S. Siau, and L. H. Wee, "A cluster analysis of EPOCH questionnaire data from university students in Sichuan, China: Exploring group differences in psychological well-being and demographic factors," Healthcare, vol. 13, no. 19, art. no. 2476, 2025, doi: 10.3390/healthcare13192476.
[47] Y. Liu, "Analysis and prediction of college students' mental health based on K-means clustering algorithm," Appl. Math. Nonlinear Sci., vol. 7, no. 1, pp. 501–512, 2022, doi: 10.2478/amns.2021.1.00099.
[48] A. H. Chowdhury, D. Rad, and M. S. Rahman, "Predicting anxiety, depression, and insomnia among Bangladeshi university students using tree-based machine learning models," Health Sci. Rep., vol. 7, no. 4, art. no. e2037, 2024, doi: 10.1002/hsr2.2037.
[49] F. Liu et al., "Use of latent profile analysis and k-means clustering to identify student anxiety profiles," BMC Psychiatry, vol. 22, art. no. 12, 2022, doi: 10.1186/s12888-021-03648-7.
[50] P. Alluri et al., "Fairness-aware K-means clustering in digital mental health for higher education students: A generalizable framework for equitable clustering," JAMIA Open, vol. 9, no. 1, art. no. ooaf174, 2026, doi: 10.1093/jamiaopen/ooaf174.
[51] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. Math. Statist. Probab., vol. 1, 1967, pp. 281–297.
[52] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowl. Discovery Data Mining (KDD), 1996, pp. 226–231.
[53] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987, doi: 10.1016/0377-0427(87)90125-7.
[54] C. X. Gao et al., "An overview of clustering methods with guidelines for application in mental health research," Psychiatry Res., vol. 327, art. no. 115265, 2023, doi: 10.1016/j.psychres.2023.115265.
[55] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Off. J. Eur. Union, L series, Jul. 2024.
[56] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, arXiv:1702.08608.
[57] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nat. Mach. Intell., vol. 1, pp. 206–215, 2019, doi: 10.1038/s42256-019-0048-x.
[58] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2017, pp. 4765–4774.
[59] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[60] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[61] H. Byeon, "Advances in machine learning and explainable artificial intelligence for depression prediction," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 6, pp. 520–526, 2023, doi: 10.14569/IJACSA.2023.0140656.
[62] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
[63] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS), 2017, pp. 1273–1282.
[64] T. B. Brown et al., "Language models are few-shot learners," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), vol. 33, 2020, pp. 1877–1901.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)