

Checklist

1. For all authors...

- (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [\[Yes\]](#)
- (b) Did you describe the limitations of your work? [\[Yes\]](#) See Section 6
- (c) Did you discuss any potential negative societal impacts of your work? [\[Yes\]](#) See Section 6
- (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [\[Yes\]](#)

2. If you are including theoretical results...

- (a) Did you state the full set of assumptions of all theoretical results? [\[N/A\]](#)
- (b) Did you include complete proofs of all theoretical results? [\[N/A\]](#)

3. If you ran experiments (e.g. for benchmarks)...

- (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [\[Yes\]](#) See section 6
- (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [\[Yes\]](#) See section 4
- (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [\[Yes\]](#) See Tables 3
- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [\[Yes\]](#) See end of first paragraph of Section 4.2

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

- (a) If your work uses existing assets, did you cite the creators? [\[N/A\]](#)
- (b) Did you mention the license of the assets? [\[N/A\]](#)
- (c) Did you include any new assets either in the supplemental material or as a URL? [\[Yes\]](#) See Sections 6 and A.2
- (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [\[Yes\]](#) See Sections 3.1 and 3.2 as well as and Figure 1
- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [\[Yes\]](#) See Sections 3.3 and 6

5. If you used crowdsourcing or conducted research with human subjects...

- (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [\[Yes\]](#) See first screenshot in Figure 1
- (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [\[Yes\]](#) See our IRB approved participant instructions which include risk in first screenshot in Figure 1
- (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [\[Yes\]](#) See Section 3.2

A Appendix

A.1 StudentSADD data description

The Student Suicidal Ideation and Depression Detection (StudentSADD) dataset includes 345 sessions of data from 302 unique student participants with modalities including participant demographics, PHQ-9 scores, responses to a text prompt, unscripted audio recordings, scripted audio recordings, and willingness to share passive modalities (Twitter, calendar, call logs, text logs, contacts, and GPS). Participants were given access to an Android and Website App. They could choose which version of the app to take. All questions were the same, however, Twitter was only passive modality assessed for willingness to share by the web app. All publicly available data is stored in CSV files.

Responses to the text prompt, demographic questions, and phq-9 questions as well as willingness to share information and timestamps of submission by participants can be found in summaryStudentSADD.csv. Participants were asked to share their favorite place (text prompt); demographic info, specifically, age, gender, student status, race/ethnicity, whether or not they have had prior depression treatment, whether they were remote, and whether or not they had covid; responses to 9 depression screening questions with four point Likert scales; and GPS data (if shared). Participant willingness to share each active prompt, each smartphone log modality, GPS, and Twitter was also included in this file. The sessions from the same student are marked as copies in this CSV file.

As the raw audio poses privacy concerns, we instead share features and embedding representations of the audio data. For example, the files scriptedSMILE.csv and unscriptedSMILE.csv contain examples of extracted openSmile audio features. These CSV files also have the PHQ-9 and item-9 scores of the corresponding participants. The VGGish embeddings for each audio recording are in the zipped folders named VGGishEmbeddingsUnscriptedStudentSADD and VGGishEmbeddingsScriptedStudentSADD. Further feature and embedding representations of the audio data will be included as part of the dataset in the future. Additionally, the file scriptedTranscript.csv contains scripted audio transcripts and the file unscriptedTranscript.csv contains unscripted audio transcripts.

For the unscriptedTranscripts and text prompt replies, proper nouns that posed privacy risks have been manually redacted. This was particularly important in cases where students named family members or friends. In no cases does this change the meaning of the reply and therefore would not impact the screening potential of the data.

Lastly, we release CSV files with results for the parameter search and baseline machine learning models in this paper. The naming convention for these files is [parameter][data][label] for the parameter search and [data][features][label] for the baseline screening results. Further, detailed results for machine learning models that use the StudentSADD data will also be released.

84 IDs were designated as a test set to evaluate the machine learning models trained on this data. The Test and Train IDs can be found below. In addition to the training IDs from unique students, we have also included the IDs from duplicated entries. Note, not all IDs shared every modality.

Test IDs: [1607712777, 292, 2613, 1610640355, 1607494599, 1607040811, 1608492986, 6390, 396, 1607734901, 1607350992, 1608992344, 1609903202, 74, 7159, 4698, 7547, 4441, 1607097951, 8479, 8170, 4707, 7516, 1609174124, 1608853150, 8516, 1611424664, 2843, 1607040596, 1953, 1607772081, 1608564004, 2627, 1607217921, 1607118643, 1607314413, 1609887404, 1608335387, 4098, 1607046006, 1608242917, 8918, 1607131299, 9754, 1607262842, 1607273026, 2478, 1607536408, 1607291545, 1608707232, 1609941585, 1608200497, 1610630377, 7711, 1607810287, 9934, 1608850448, 4041, 1609166629, 1608168856, 1607572897, 6831, 1608586814, 1608588581, 2837, 8180, 1608631410, 1607051003, 3830, 4879, 1608920128, 1607019351, 8181, 3473, 1608335906, 1607738757, 1608770486, 7564, 1607495239, 1609983150, 1607397061, 1607696074, 103, 2222]

Unique Train IDs: [4769, 1607928177, 1607269923, 7755, 4598, 1607807806, 1608741452, 3323, 1610110670, 1607133044, 9745, 1607291670, 5245, 4442, 319, 1607133218, 1607010270, 1608587203, 1609256130, 1608582258, 5028, 1609771771, 5229, 3517, 1608595561, 1608048050, 1607410780, 528, 1607134906, 3102, 1607555727, 1609887167, 3985, 7256, 3523, 1607289708, 1609890222, 850, 1608917024, 5047, 1608061691, 4782, 1608062276, 1056, 1611517276, 1607636681, 1607891972, 5571, 1609052616, 1607927243, 2525, 4353, 1610818662, 8640, 1607559849, 6706, 1608624428, 1607968838, 1608672132, 552, 1608537399, 1610381937, 1608607986, 381, 1608589576, 3920, 1608059746, 1609027319, 1607357022, 1607691623,

574 1609899907, 1608470962, 8791, 1610380419, 3064, 1609473849, 1607712704, 1609887249,
575 1609888813, 1608588103, 1244, 7279, 1607339125, 1607712682, 8472, 1269, 1607045076,
576 1607365865, 1846, 191, 1811, 1608702785, 1609049435, 5330, 1607257348, 1609890530,
577 3278, 1608586899, 1607939718, 2430, 1609893292, 60, 1607270186, 6336, 8650, 1608495626,
578 1608586953, 2121, 1607295286, 896, 1609889389, 1607560754, 6548, 6580, 1607440988,
579 1609111416, 1607807159, 8663, 1607129044, 6658, 1607799213, 3933, 1608596696, 1608663032,
580 1610791060, 1607135820, 1607413039, 1607659758, 1608487726, 4859, 1609142183, 1607276888,
581 7452, 1607368510, 1607266081, 2623, 1608416516, 2128, 3227, 5881, 6510, 1609166843,
582 7569, 1607712793, 1608850996, 3273, 1607939838, 9986, 3302, 1607206195, 1609082904,
583 1607510222, 7612, 1607022963, 1607051040, 1607719324, 1608849324, 1607642639, 1607104225,
584 705, 1608506424, 1608188073, 8018, 8085, 4755, 1611704179, 1607193886, 7007, 3041, 4001,
585 1552, 1716, 1608053349, 1608572299, 1608051417, 1607712784, 836, 1607929944, 1607795480,
586 1608200317, 415, 3662, 1610109929, 2496, 8550, 6868, 1608587385, 1608591490, 7370, 4549,
587 7505, 1879, 1876, 1608003341]

588 Duplicated Train IDs: [1607315588, 1607087749, 518, 6280, 1607315467, 1607497867,
589 1608359052, 8468, 1607124379, 6941, 1607510942, 1607921053, 1611575587, 1608201126,
590 1607639591, 4521, 1607315498, 1607639340, 1607785646, 1609353141, 1607453496, 1607293881,
591 1607674683, 5948, 1608683584, 1608087234, 3267, 1607802179, 7109, 838, 9034, 1868,
592 1608683738, 8284, 95, 1609154655, 1607159523, 1607124838, 1607088999, 7912, 1607802347,
593 1607764720, 1607785968]

594 **A.2 Data and code access for reviewers**

595 To access the StudentSADD data, reviewers can use the login information (provided in the submission
596 system) at our project website: <https://emutivo.wpi.edu/data/>

597 Upon publication, we will grant similar access to other researchers who request our data and agree to
598 the terms of the data licence in Section A.3. This approach is standard for datasets in this domain
599 [31; 33]. Data access will be managed by the Academic and Research Computing team at WPI.

600 This is not a static dataset, as we will continue to add more data features and embedding representa-
601 tions. We have inquired about obtaining a DOI for StudentSADD through our institutional library,
602 and will obtain a DOI prior to publication.

603 To access the code used to generate the results in this paper (including the parameter search), reviewers
604 can navigate to our public github repository: <https://github.com/mltlachac/StudentSADD>

605 **A.3 StudentSADD Dataset - End User License Agreement**

606 (<https://emutivo.wpi.edu>)

607 The person in request may download and use this database only after signing and returning this
608 agreement form. By signing this document, the user agrees to the following terms:

609 **Commercial and academic use**

610 The database is made available for research purposes only. Any commercial use of this data is
611 forbidden.

612 **Redistribution**

613 The user may not distribute the database or parts of it to any third party.

614 **Publications**

615 The use of data for illustrative purposes in publications is allowed. Publications include both scientific
616 papers and presentations for scientific/educational purposes. In this case, the identity of the subjects
617 should be protected (no release of identifiable information for subjects).

618 **Citation**

619 All publications reporting on research using this database have to acknowledge this by citing the
620 following article:

621 *ML Tlachac, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican,*
622 *Connor Bruneau, Hunter Caouette, Eral Toto, Elke Rundensteiner, "StudentSADD: Mobile De-*
623 *pression and Suicidal Ideation Screening of College Students during the Coronavirus Pandemic",*
624 *in Submission at Neural Information Processing System (NeurIPS) 2021 Datasets and Benchmarks*
625 *Track*

626 For specific software output that is shared as part of this data, the user agrees to respect the individual
627 software licenses and use the appropriate citations as mentioned in the documentation of the data.

628 **EULA changes**

629 Worcester Polytechnic Institute is allowed to change these terms of use at any time. In this case,
630 users will be informed of the changes and will have to sign a new agreement form to keep using the
631 database.

632 **Warranty**

633 The database comes without any warranty. In no event shall the provider be held responsible for any
634 loss or damage caused by the use of this data.

635

636 _____

637 Name Date Signature

638

639 _____

640 Organization Address

641 **A.4 Data management plan**

642 **Data description and formats**

643 The Student Suicidal Ideation and Depression Detection (StudentSADD) dataset contains 245 sessions
644 submitted through a mobile app by 302 students over between August 2020 and January 2021. The
645 data includes PHQ-9, demographics, retrospective phone data, text prompt, audio recordings, location
646 history, and tweets. The audio is stored as WAV files. The remaining data is stored as text in CSV
647 files.

648 **Data archiving, access and sharing, and data preservation**

649 Data will be stored in the file systems of Worcester Polytechnic Institute (WPI). The stored data will
650 be protected with disk mirroring, daily backups and other means. Full-time system administrators will
651 monitor the security and availability of these systems. Appropriate access control and other security
652 policies and mechanisms will be put in place to protect the integrity, security, privacy, confidentiality
653 and other rights or requirements. For access by and sharing with the greater research community
654 and general public, research group websites will be used, and personally identifiable information
655 (PII) will be appropriately removed or anonymized from the essential data used for our research. To
656 protect privacy, confidentiality or other rights/requirements, while at the same time ensuring scientific
657 reproducibility and verifiability, the data will be summarized in the forms of intermediate statistics,
658 and made publicly available, or when requested from other researchers.

659 **Data privacy**

660 Data will be shared only under rules specified by our IRBs, and only when properly anonymized.
661 The systems on which the data will be stored will have access restricted to the project members via
662 standard filesystem permissions management. The server on which the data is stored is within the
663 University's restricted access datacenter. Files that will be made publicly available will not contain
664 any identifiable location information.

665 **Policies and provisions for re-use, re-distribution**

666 Any data collected will be in compliance with the IRB protocols of Worcester Polytechnic Institute.
667 Data gathered for this project may be reused in other, related, research projects conducted by the PIs
668 or graduate students. It is expected that other researchers in machine learning in mental health would
669 be interested in our dataset. Requests for the data would be handled as per Section III. These datasets
670 cannot be used for commercial applications or purposes, or changed and resubmitted without the PIs'
671 permission and are subject to the intellectual property policies of WPI.

672 **Rights and Obligations**

673 The Principal Investigators (PIs) will be responsible for the implementation of the Data Manage-
674 ment Plan. WPI owns the technology developed at each individual university by each university's
675 employees.