Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones

ARVIND PILLAI, Dartmouth College, USA
SUBIGYA NEPAL, Dartmouth College, USA
WEICHEN WANG, Dartmouth College, USA
MATTHEW NEMESURE, Dartmouth College, USA
MICHAEL HEINZ, Dartmouth College, USA
GEORGE PRICE, Dartmouth College, USA
DAMIEN LEKKAS, Dartmouth College, USA
AMANDA COLLINS, Dartmouth College, USA
TESS GRIFFIN, Dartmouth College, USA
BENJAMIN BUCK, University of Washington, USA
SARAH MASUD PREUM, Dartmouth College, USA
TREVOR COHEN, University of Washington, USA
NICHOLAS JACOBSON, Dartmouth College, USA
DROR BEN-ZEEV, University of Washington, USA
ANDREW CAMPBELL, Dartmouth College, USA

Speech-based diaries from mobile phones can capture paralinguistic patterns that help detect mental illness symptoms such as suicidal ideation. However, previous studies have primarily evaluated machine learning models on a single dataset, making their performance unknown under distribution shifts. In this paper, we investigate the generalizability of speech-based suicidal ideation detection using mobile phones through cross-dataset experiments using four datasets with N=786 individuals experiencing major depressive disorder, auditory verbal hallucinations, persecutory thoughts, and students with suicidal thoughts. Our results show that machine and deep learning methods generalize poorly in many cases. Thus, we evaluate unsupervised domain adaptation (UDA) and semi-supervised domain adaptation (SSDA) to mitigate performance decreases owing to distribution shifts. While SSDA approaches showed superior performance, they are often ineffective, requiring large target datasets with limited labels for adversarial and contrastive training. Therefore, we propose *sinusoidal similarity sub-sampling (S3)*, a method that selects optimal source subsets for the target domain by computing pair-wise scores using sinusoids. Compared to prior approaches, S3 does not use labeled target data or transform features. Fine-tuning using S3 improves the cross-dataset performance of deep models across the datasets, thus having implications in ubiquitous technology, mental health, and machine learning.

Authors' addresses: Arvind Pillai, arvind.pillai.gr@dartmouth.edu, Dartmouth College, 15 Thayer Dr, Hanover, New Hampshire, USA, 03755; Subigya Nepal, Dartmouth College, Hanover, USA; Weichen Wang, Dartmouth College, Hanover, USA; Matthew Nemesure, Dartmouth College, Hanover, USA; Michael Heinz, Dartmouth College, Hanover, USA; George Price, Dartmouth College, Hanover, USA; Damien Lekkas, Dartmouth College, Hanover, USA; Amanda Collins, Dartmouth College, Hanover, USA; Tess Griffin, Dartmouth College, Hanover, USA; Benjamin Buck, University of Washington, Seattle, USA; Sarah Masud Preum, Dartmouth College, Hanover, USA; Trevor Cohen, University of Washington, Seattle, USA; Nicholas Jacobson, Dartmouth College, Hanover, USA; Dror Ben-Zeev, University of Washington, Seattle, USA; Andrew Campbell, Dartmouth College, Hanover, USA.

2474-9567/2023/12-ART174
https://doi.org/10.1145/3631452

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Computing methodologies → Learning under covariate shift**; **Transfer learning**; • **Applied computing → Health informatics**.

Additional Key Words and Phrases: speech, domain adaptation, domain generalization, suicidal ideation, cross-dataset, mobile phones, smartphones

ACM Reference Format:
Arvind Pillai, Subigya Nepal, Weichen Wang, Matthew Nemesure, Michael Heinz, George Price, Damien Lekkas, Amanda Collins, Tess Griffin, Benjamin Buck, Sarah Masud Preum, Trevor Cohen, Nicholas Jacobson, Dror Ben-Zeev, and Andrew Campbell. 2023. Investigating Generalizability of Speech-based Suicidal Ideation Detection Using Mobile Phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 174 (December 2023), 39 pages. https://doi.org/10.1145/3631452

1 INTRODUCTION

Suicidal Ideation (SI) is a significant public health concern that affects an estimated 12.4 million adults (5.0% of the US population) according to the National Survey on Drug Use and Health (NSDUH) [76]. SI refers to thoughts or contemplation of suicide. It is a serious mental health concern and a precursor to suicidal behavior [44]. The proliferation of ubiquitous technology has led to advanced screening methods for SI. Speech-based mobile systems are one such method that collects audio recordings through smartphones and analyzes paralinguistic patterns using machine learning (ML) [97, 104].

However, previous research indicate that ML methods exhibit performance decreases under distribution shifts [81], raising serious safety concerns. In fact, most prior speech-based SI screening methods are only evaluated on a single dataset or specific populations owing to the high costs associated with conducting longitudinal studies in mental health, and privacy-related barriers to sharing data across institutions. It is imperative to understand the generalization capabilities of these methods for smooth deployment. Consequently, a well-known journal recently established best practices for implementing machine learning in healthcare, highlighting limited generalizability of models and their tendency to exacerbate biases in data [53], while a leading digital health journal emphasized the importance of independent validation [14]. Challenges in speech generalization for mental health arise from several sources. First, the incidence of SI varies depending on the population. For example, a clinical population reporting symptoms of persecutory ideation will have more individuals with SI than student populations [35]. Such differences can be observed across different clinical populations. Second, the machine learning methods used for pattern recognition might be highly sensitive to population characteristics and small datasets. Balancing intra-dataset and inter-dataset performance remains a fundamental challenge. Third, audio characteristics may vary across samples.

Consequently, the effect of these out-of-distribution shifts in speech-based SI detection is largely unknown to the research community, which is a significant gap. Thus, we sought to examine SI detection performance across four datasets. To this end, we collect three datasets investigating a mental illness or related symptoms such as major depressive disorder, auditory verbal hallucinations, and persecutory ideation. In addition, we include the open-source StudentSADD dataset [104] for analysis. Therefore, we evaluate generalizability across four datasets for speech-based SI detection. We first employ a consistent strategy to select data and extract features that allow fair evaluation. Subsequently, we examine the out-of-distribution generalization across the four datasets with a binary classification task – using speech to classify whether an individual has SI.

We systematically evaluate cross-dataset performance by first investigating dataset similarity qualitatively (t-SNE visualizations) and quantitatively (OTDD metric). Next, we assess within-dataset performance using machine learning (ML) and deep learning (DL) models in a stratified-k-fold setup. After establishing strong baselines, we examine the performance of models when trained on one dataset and tested on another, referred to as one-one validation. Similarly, we experimented with leave-one-dataset-out validation, which completely holds out one dataset for testing, and trains with the rest. Our results indicated poor generalization performance of ML and DL methods and emphasizes choosing optimal data points for training. Thus, we applied UDA and SSDA approaches to improve generalization. While UDA methods apply feature transformations and instance weighting, SSDA methods rely on using limited target label data in a contrastive or adversarial training setup. Both assume access to large unlabeled target datasets, making them less ideal for mental health.

In this paper, we propose an SSDA method termed sinusoidal similarity sub-sampling (S3) that works with smaller unlabeled target datasets. S3 selects an optimal subset from the source dataset to adapt to the target dataset, and it is computed as follows. First, we transform the source and target domain embeddings from a deep learning model (VGGish [45]) into sinusoidal signals. Next, we randomly generate an *anchor* sinusoidal matrix composed of many sine waves. Finally, the transformed embeddings are compared to the *anchor* through dot products to obtain pair-wise scores to select best source subset in different ways, which are referred to as S3 variants. Intuitively, S3 extracts frequency information by comparing a series of sine waves. Using the subset for fine-tuning models results in better generalization performance than other methods.

1.1 Contributions

To investigate the generalizability of speech-based SI detection methods, our contributions are as follows:

- We evaluate the performance of SI detection across four different datasets with three experiments: within-dataset (section 5.2), one-one (section 5.3), and leave-one-dataset-out (section 5.4) validation. Consequently, we benchmark several UDA and SSDA methods based on their effectiveness in handling the dataset shift. In general, our results indicate SSDA methods performed better than UDA approaches. To our knowledge, we are the first to validate speech-based SI detection on multiple independent datasets, elucidating previously unknown factors.
- We propose the sinusoidal similarity sub-sampling (S3) metric with a focus on improving generalization in the context of mental health, where target datasets are small and unlabeled. We observe that S3 outperforms UDA and SSDA methods in many cross-dataset scenarios. S3 obtained significant performance gains for the smallest dataset (n<50).
- We perform extensive post-hoc analysis to interpret important features for generalization across different populations. Our findings suggest spectral roll-off is crucial across two datasets but not the other, suggesting some acoustic heterogeneity may exist across datasets even among commonalities. Furthermore, we evaluate the robustness of UDA and SSDA approaches to help future researchers choose appropriate methods. Finally, our analysis on using S3 for acoustic scene classification indicates that it is well-suited for mental health, but not as a general audio SSDA method.

2 BACKGROUND & RELATED WORK

2.1 Ubiquitous sensing and mental health

Passive sensing data from smartphones and wearables show strong potential for identifying individuals with mental illness [47, 70, 114]. StudentLife [113] was the first passive sensing Android application to assess mental health. Recently, several studies have employed the use of smartphones and mobile applications for depression detection [72, 114]. For instance, Mullick et al. [72] collected data from 55 adolescents to predict depression; their findings highlight the utility of screen, call, and location-based features in improving performance. A study by Xu et al. [114] investigated the generalizability of various sensing data across different populations for depression detection, suggesting the need for improved methods that can be validated across independent datasets. Although several studies have addressed the problem of depression detection, SI has received limited attention. Horwitz et al. [47] investigated the prediction of suicidal ideation (SI) among medical interns using FitBit data on sleep and steps. They found that passively collected FitBit data did not enhance SI detection. They also acknowledged that better results were achieved when data collection was closer to the outcome rather than averaging sensing data over time. Sleep serves as a crucial biomarker for mental health, as demonstrated by Wang et al. [112], who discovered a connection between delayed bedtimes and self-reported concerns of potential harm and hallucinations in individuals with schizophrenia. Additionally, Abdullah et al. [1] explored circadian rhythm through sleep period markers to detect sleep deprivation and enhance overall well-being. To detect bipolar symptoms, Gruenerbl et al. [43] validated accelerometer-location sensors with bipolar patients from an Austrian psychiatric hospital. Their system achieved 72-81% accuracy in recognizing clinical states (depression/mania) and demonstrated high precision (96%) and recall (94%) in state-change detection.

2.2 Suicidal ideation detection

Various methods such as electronic health records (EHR) [7, 111], functional MRI (fMRI) [54, 64, 73], video [60, 94], social media text [26, 27, 49, 77], and speech [9, 16, 22, 24, 104] can detect suicidal ideation. Rich longitudinal EHR data with information on diagnostic codes, laboratory results, and medications are particularly useful. For example, Barak-Corren et al. [7] used a Bayesian model with 15 years of EHR data to predict future suicidal behavior, observing that unconventional factors like back contusions can increase suicide risk. Similarly, Walsh et al. [111] used EHR data from 5167 participants to predict suicide attempts on a larger scale. They applied a random forest to predict suicide attempts within a seven-day window, achieving a 0.84 AUC. In brain imaging, Li et al. [64] found that voxel-wise concordance in parts of the brain can be used as a biomarker for SI in individuals with depression. Another study by Nawaz et al. [73] concluded that there was no evidence to support the association between SI and amygdala structural changes. Videos are another useful tool to analyze body and facial cues, and thus enable SI detection. For instance, Shah et al. [94] examined social media videos suggesting that multi-modal information combined with shoulder and torso changes are important features for SI detection. Another study by Laksana et al. [60] examined facial behaviors and observed that smile-based descriptors are the most discriminative for SI detection

Social networking websites such as Facebook, Twitter, and Reddit provide an anonymous space for individuals to share their suicidal thoughts. Many studies have sought to detect SI using text data scraped from these websites [26, 27, 49, 77]. For example, O'dea et al. [77] extracted 14,701 tweets from Twitter and trained an SVM to classify highly concerning tweets automatically. Reddit is perhaps the most relevant platform for studying SI [26, 27, 119]. De Choudhury et al. [27] sought to forecast if a person talking about mental health online would transition into suicidal ideation discussions on Reddit. A similar study by De Choudhury and De [26] investigated self-disclosure, anonymity, and social support on Reddit mental health forums. Their results suggest that responses are surprisingly high quality and contain prescriptive advice, contrasting responses on Twitter. Text messages are a useful modality for SI detection. While Nobles et al. [75] address the subtle problem of differentiating between suicidal and depression periods, Tlachac et al. [103] detect SI using less longitudinal data, i.e., predicting a particular week's SI using data from previous weeks.

Compared to many methods mentioned above, mobile phones enrich longitudinal diary studies and psychology research with their in-the-wild data collection capabilities [96]. Experience Sampling Methods (ESM) use these devices to trigger prompts for user data, providing researchers with timely information, reducing recall bias [59, 62]. Speech-based methods have several advantages over traditional approaches, they are easy to use, relay real-time longitudinal data to researchers, and facilitate sharing of personal narratives [90]. Furthermore, audio diaries are discreet, time-saving, and capture authentic emotions. Their convenience promotes user compliance, allowing more open sharing of sensitive information [15, 46].

Speech has emerged as an important active modality for SI screening, where machine learning models are used to learn patterns from paralinguistic features. In particular, the AVEC2013 [107] feature set is widely used for depression and SI screening [23, 24, 104]. Broadly, studies can be classified as those using data from clinical interviews [16] with long recordings, or smartphones with shorter in-situ recordings [9, 104]. A study by

Table 1. An overview of suicidal ideation studies using speech.

Study	Participants	Mobile/Scalable	Independent dataset validation	Ground truth	Population Characteristics
Chakravarthula et al. [16]	124	✗	✗	Interview	Non-clinical
Stasak et al. [97]	246	✗	✗	Hospital records	Mixed
Gideon et al. [40]	31	✓	✗	Hospital records	Mixed
Belouali et al. [9]	124	✓	✗	PHQ-9	Mixed
Tlachac et al. [104]	302	✓	✗	PHQ-9	Non-clinical
Ours	786	✓	✓	PHQ-9	Clinical, Non-clinical, Mixed

Chakravarthula et al. [16] examined suicide risk factors in military couples with acoustic features, embeddings, and lexical cues. Using these features, they trained an SVM to predict categories of suicidal risk such as none, ideation, attempt with an average recall of 0.6. Similarly, Stasak et al. [97] investigated manually annotated voice quality and speech disfluency measures in 246 individuals with and without SI. Their findings suggest that the SI group has a lesser average number of hesitations and speech errors compared to the suicide attempt group.

Belouali et al. [9] investigated SI detection in veterans using voice recording from Android tablets. They train machine learning models on phonation, prosody, and glottal features of voice to obtain an AUC of 0.78. Another study [40] modelled emotions such as guilt and anger from phone conversations to detect SI obtaining an AUC of 0.79. However, their dataset had only 31 participants. More recently, a study by Tlachac et al. [104] investigated speech patterns in over 300 students for SI detection. While their traditional machine learning methods used the AVEC2013 feature set, their deep learning model trained on unscripted audio obtained a balanced accuracy of 0.73. Most speech-based SI detection studies are evaluated on a single dataset or population, making them prone to poor cross-dataset generalization and data biases [14, 53, 81]. In contrast, we sought to understand performance across four datasets. Thus, our investigating differs from the previous studies in the following ways. First, to our knowledge, we are the only study to validate speech-based SI detection on multiple independent datasets (Table 1). To this end, we analyze four datasets comprising clinical, non-clinical, and mixed populations. Second, our analysis of 786 participants is larger than previous studies, with highly varying positive SI samples ranging from 11% to 74%. Third, our studies use audio diaries collected from our Android application with the same core system, establishing a paradigm for scalable data collection. Identifying SI in near real-time is crucial to administering interventions. Mobile applications are better than high-burden clinical interviews in this regard. Furthermore, smartphones enable in-situ data collection, thus, reducing costs and improving diversity by enrolling individuals from underrepresented communities. Audio diaries have some advantages over social media content analysis - they may be accompanied by contemporaneously collected ground truth, such as item 9 from the PHQ-9 [57, 58].

2.3 Domain adaptation

Evaluating models trained on a specific population against a different population under distribution shift is crucial for real-world deployment of speech-based mental health screening systems. The poor generalization performance in such scenarios can be alleviated through Domain adaptation (DA) [29]. Given a source domain $\mathcal{D}_S$ and a target domain $\mathcal{D}_T$ with source and target joint probability distributions P^S_{XY} and P^T_{XY}, respectively. DA assumes distribution shifts where $P^S_{XY} \neq P^T_{XY}$.

Unsupervised domain adaptation (UDA) is well studied, and many methods have been proposed for computer vision (CV) and natural language processing (NLP) [33, 39, 100, 105]. In UDA, we have a labeled source dataset

$\mathcal{D}_S = \{(x_s, y_s)\}$ and a large unlabeled target dataset $\mathcal{D}_T = \{(x_t)\}$. Traditional feature-based DA techniques like subspace alignment [33], and transfer component analysis [80] transform features such that the latent spaces of the source and target domains are closer [33, 80]. For deep learning models, adversarial domain adaptation has been widely studied [4, 39, 105, 117]. These methods generally seek to build new feature representations for source and target data, making them indistinguishable for a discriminator network to classify. In instance-based DA methods such as linear discrepancy minimization [67] and kernel mean matching [42], source data is re-weighted to minimize the distance between source and target joint distributions.

In semi-supervised domain adaptation (SSDA), we have a labeled source dataset $\mathcal{D}_S = \{(x_s, y_s)\}$ and a small unlabeled target dataset $\mathcal{D}_T = \{(x_t, y_t)\}$, and a large unlabeled target dataset $\mathcal{D}_u = \{(x_u)\}$. Grandvalet and Bengio [41] proposed a method to adapt neural networks by minimizing the entropy on unlabeled target data, whereas Saito et al. [91] showed that using adversarial training to maximize the entropy followed by minimization improves the quality of discriminative features. Kim and Kim [55] introduce the concept of intra-domain discrepancy where target sub-distributions are unaligned and propose a three-step procedure for mitigation. Forgoing adversarial training, Singh [95] presents a contrastive learning framework that learns good representations through strongly augmenting unlabeled target data. Recently, Yu and Lin [115] proposed to denoise the source data by viewing it as a noisily-labeled version of the target data.

The performance of the above-mentioned approaches for speech and mental health remains largely unknown. Additionally, these methods assume access to large unlabeled target datasets to enable adversarial and contrastive training, which is uncommon in mental health. In contrast, our work proposes a sub-sampling method to select the most optimal source subset for fine-tuning the target dataset without the need for large datasets or target domain labels.

3 STUDY

Our analysis uses speech data from four studies that study mental illness in a specific population. We refer to these datasets as MDD, AVH, PT, and Student, referencing individuals with major depressive disorder, auditory verbal hallucinations, persecutory thoughts, and students, respectively. In this section, we describe the datasets (section 3.1), speech-based diaries (section 3.2), and ground truth (section 3.3).

3.1 Datasets

We use data from three of our studies - MDD, AVH, and PT - as well as an open-source dataset, StudentSADD. For brevity, we only discuss characteristics pertinent to this paper. We provide additional information about study protocols, collection prompts, and the Android application in the supplementary materials.

3.1.1 MDD. The MDD [74] study aims to 300 recruit participants with Major Depressive Disorder (MDD) from across the United States. This study challenges two widely held assumptions about MDD. First, current diagnostic criteria assume that MDD symptoms are interchangeable, i.e., determining whether an individual has MDD based on their total PHQ-9 score. However, this method fails to acknowledge the vast variance in MDD symptom presentations across individuals [38]. In fact, MDD has over 1000 unique symptoms [21]. Second, current diagnostics assume that MDD remains stable across weeks and even months. In contrast, MDD symptom intensity can vary substantially, even across a single day. [31, 37].

To address the above-mentioned issues, our study was designed with the goal of using passive sensing and EMA data to predict within-person changes in MDD symptoms, with the understanding that MDD is both highly variable across individuals and a changing system. Qualifying participants install our Android application for 90 days and answer three PHQ-9 surveys each day to facilitate within-day analysis of symptoms using smartphone data. After Item 9 in PHQ-9, the user can record an audio diary (see Fig. 1), resulting in a one-one mapping between audio recordings and PHQ-9 surveys. Note that recording audio diaries are completely optional, thus,

Table 2. Demographic descriptors of participants in our analysis. Note that pacific islander, hispanic/latino individuals are grouped under Other. (Demographics for one MDD participant is unavailable).

Category	**MDD (Clinical)**	**Student (Non-clinical)**	**AVH (Mixed)**	**PT (Mixed)**
Overall				
Participants	43	178	356	209
Audio Samples	43	178	356	209
Suicidal Ideation	5 (11.6%)	37 (21%)	214 (60.11%)	156 (74.6%)
Gender				
Male	4 (9.3%)	66 (37.0%)	154 (43.2%)	53 (25.3%)
Female	37 (86.04%)	104 (58.4%)	192 (53.9%)	150 (71.7%)
Transgender	N/A	N/A	8 (2.2%)	4 (1.9%)
Other	1 (2.3%)	8 (4.4%)	2 (0.5%)	2 (0.9%)
Race				
White	30 (69.8%)	115 (64.6%)	224 (62.9%)	147 (70.3%)
Black or African American	5 (11.6%)	6 (3.3%)	80 (22.4%)	42 (20.0%)
More than one race	6 (13.9%)	10 (5.6%)	38 (10.6%)	13 (6.2%)
American Indian/Alaska Native	0	0	7 (1.9%)	2 (0.9%)
Asian	1 (2.3%)	38 (21.3%)	4 (1.1%)	3 (1.9%)
Other	0	9 (5.0%)	N/A	1 (0.4%)

some participants did not choose to submit speech samples. Currently, the study is live with 182 completed participants. Participants were compensated $1 per EMA completed and bonuses for maintaining high compliance. Additional information about the MDD study can be found in [74].

3.1.2 AVH. The AVH study aims to collect mobile sensing data from individuals experiencing auditory verbal hallucinations (AVH). AVH are prevalent in people with psychiatric diagnosis and healthy individuals [61], and measuring distinguishing factors between the groups is challenging. Therefore, this study has two main contributions. Firstly, it uses the Research Domain Criteria (RDoC) [25] framework from the National Institute of Mental Health to investigate AVH on a spectrum from "normal" to pathological. Secondly, it utilizes a smartphone app to gather data through passive sensing, audio diaries, and momentary self-assessments, thus differing from traditional retrospective methods like interviews and surveys that can be prone to inaccuracies. Using the above-mentioned factors, the study aims to evaluate whether AVH experience and behavior differ across clinical and non-clinical individuals. For more details about the study, we refer the reader to [11].

We utilized the Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) self-report to evaluate AVH [109]. In total, 384 participants met the recruitment criteria. Among them, 192 were female, 176 were male, and 12 identified as transgender (male to female and female to male). Four participants identified as another gender. Participants installed our Android app on their phone that was designed to collect mobile sensing, audio diaries, and self-report ecological momentary assessments. We modified these tools to fit the needs of our study. During the 30-day study, we collected both active modalities, like audio diaries, and passive sensing data, like GPS, telemetry, and light data.

3.1.3 PT. This study [13] aims to understand persistent harmful thoughts, called persecutory thoughts (PT). PT is prevalent in various mental health conditions, including mood, anxiety, personality, and neurodegenerative disorders. It is also found in healthy individuals. Similar to AVH, research supports a continuum of PT ranging from normal thoughts about danger to strong negative beliefs that disrupt daily life. Thus, our team sought to study the phenomenology of PT. The novel contributions of this study are as follows. First, we aimed to characterize how often people experience different aspects of paranoia-related thoughts, feelings, and actions in their daily lives. Second, we evaluate the link between these aspects and levels of clinical severity defined by treatment received. Third, we explored if people with greater functional disability exhibited similar paranoia experiences. Additional study details can be found in [13].

Here, we modify the Android Application used in the AVH study to fit the needs of the PT study. We collected data from 231 individuals who experienced PT using the Revised Green Paranoid Thoughts Scale's ideas of persecution (R-GTPS) subscale [36]. The R-GTPS is a 10-item measure of persecutory ideation that was derived from the full-length Green Paranoid Thoughts Scale. Similar to previous studies, participants were recruited remotely through Google Ads. They were instructed to keep their android application installed smartphones with them throughout the 30-day data collection period and complete brief questionnaires as prompted. They could contact a research coordinator for technical support if needed, and the research team would follow up if no information was received from their devices for three days. Participants received $125 as compensation for their participation, and the app was uninstalled at the end of the data collection period.

3.1.4 Student.

The StudentSADD [104] dataset (Student) aimed to study suicidal ideation and depression among students during COVID-19. For data collection, two aesthetically similar Android and Web applications were developed. The app collected PHQ-9 depression screening surveys, demographics, a typed reply, two voice recording, and Twitter information from all participants. In total, 302 participants submitted their sessions. Detailed information about the StudentSADD study is described in [104]. The dataset consists of several feature sets extracted from raw audio for analysis. To protect participant's personal identifiable information, raw audio information was not released by the authors. Note that we only use unscripted audio features for analysis because it closely matches data in other studies. Therefore, we use 178 audio recordings with SI labels (see Table 2). Moreover, we do not use the Student dataset for deep learning analysis as it requires raw audio data.

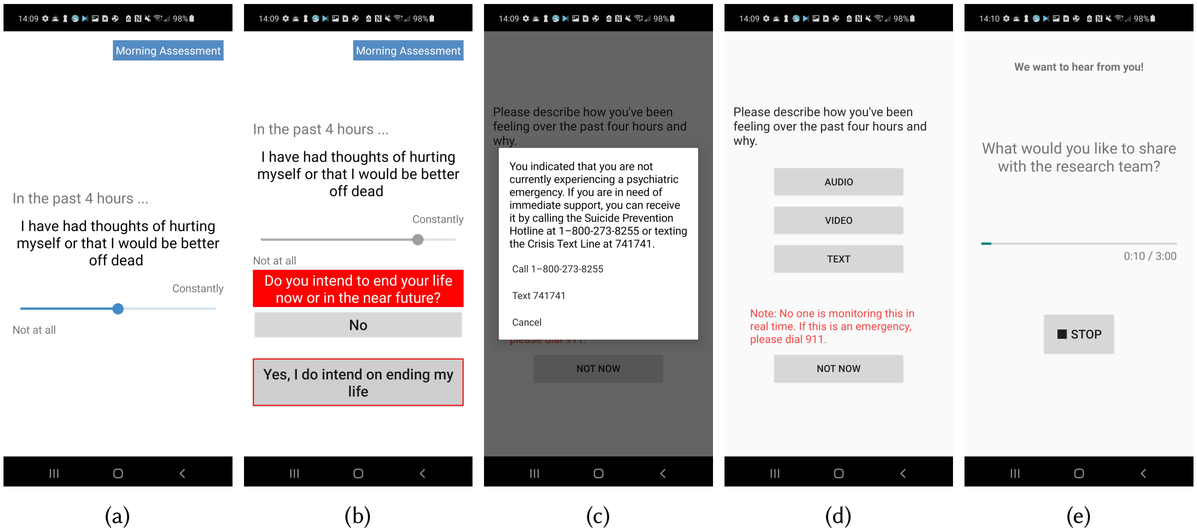
3.2 Speech-Based Diaries

The Android application used in the MDD, AVH, and PT studies is inspired by mobile sensing systems used in other mental health and behavioral studies [10, 113]. In addition, they have been tailored to suit the specific requirements of each study. To ensure strong adherence to the study, we utilize the subsequent techniques for designing, integrating, and deploying the app. The app is designed to function in the background of mobile phones, automatically collect passive sensing data, and prompt the user to collect active modalities such as audio diaries through EMAs. The voice recordings use similar protocols. First, the diaries can be atmost 180 seconds and are completely optional, thus, reducing the likelihood of trivial submissions. Second, the prompts used for collection are unrelated to SI, consequently facilitating analysis of low-level paralinguistic features. Third, speech is unconstrained, i.e., we do not instruct the user to repeat a sentence or read a paragraph. Fourth, it is collected in-situ without any restrictions on the participant's location, resulting heterogeneous speech signals. To summarize, we collect free-form speech from the participant's phone under naturalistic conditions. We envision that detection systems tested in in-the-wild noisy conditions can catalyze the development of robust SI intervention protocols.

While the core system remained the same across study, some factors are tailored towards studying the primary population group. The groundtruth collection for each study is as follows:

MDD. The app prompts the user three times a day with the PHQ-9 questionnaire. After the Item 9, the user can record an audio diary as shown in Fig. 1.

AVH. To record an audio diary, first, the HPSVQ [109] is administered four times a day for 30 days, randomly between 9am-12pm, 12pm-3pm, 3pm-6pm, and 6pm-9pm. Note that the PHQ-9 is collected only once at baseline during on-boarding.



In the past 4 hours ...

I have had thoughts of hurting myself or that I would be better off dead

Do you intend to end your life now or in the near future?

Yes, I do intend on ending my life

Please describe how you've been feeling over the past four hours and why.

You indicated that you are not currently experiencing a psychiatric emergency. If you are in need of immediate support, you can receive it by calling the Suicide Prevention Hotline at 1–800-273-8255 or texting the Crisis Text Line at 741741.

Call 1–800-273-8255

Text 741741

Cancel

Note: No one is monitoring this in real time. If this is an emergency, please dial 911.

We want to hear from you!

What would you like to share with the research team?

Fig. 1. Example Android application screens from the MDD study: (a) The PHQ-9 Item 9 question, (b) The user submits a high Item 9 score and is redirected to a safety question which provides immediate support, (c) A direct link to call emergency services if the user is currently experiencing SI, (d) The user can optionally submit a audio, video, or text diary, (e) The audio diary screen where the user can submit a recording up to 180 seconds. The safety protocols employed are described in section 7.4.

PT. The user is prompted four times a day semi randomly between 9am-9pm to answer a 12 item survey measuring PT, cognitive appraisals, anxiety, self-esteem, sadness, sociality, energy, and presence of others. Note that the PHQ-9 is collected only once at baseline during on-boarding.

Student. In the StudentSADD study [104], the app prompts the user with a general question such as "Describe a good friend". The user has 30 seconds to record their unscripted voice sample.

Additional information regarding studies, android application, and audio diaries are presented in the supplementary.

3.3 Ground truth

The Patient Health Questionnaire (PHQ-9) is a commonly used self-report tool for measuring the severity of depression with high validity and reliability [57, 58]. The survey consists of nine items that the participant rates on a Likert scale ranging from 0-3. The responses correspond to "not at all," "several days," "more than half the days," and "nearly every day," respectively. Item 9 in the PHQ-9 screens for suicidal ideation, and it is a strong predictor of suicide attempts [89]. It asks how often the individual has been bothered by "thoughts that you would be better off dead, or thoughts of hurting yourself in some way?" Any answer other than 0 or "not at all" is considered SI. In the MDD study (section 3.1.1), the PHQ-9 scale was modified to range from 0-100, and any value over 24 is considered SI.

For fair evaluation, we ensure each participant has only one audio sample (1:1). As data collection frequency differs across studies, we describe the selection criteria here. All unscripted audio recordings with SI labels are used in the student dataset. We use the following selection criteria for the MDD, AVH, and PT datasets: (1) Audio samples must have a voice, measured using $F_0 > 27.5$Hz and word count greater than 0, (2) the audio length must be at least 30 seconds. In the MDD dataset, we select the longest audio recording the participant has submitted.

In the AVH and PT datasets, we choose the longest audio samples submitted in the first two days of the study. Demographic information and statistics of the dataset used for analysis are presented in Table 2.

4 METHODS

4.1 Dataset Similarity

Prior to evaluating predictive performance across datasets, it is crucial to understand their similarities. Thus, we use t-distributed stochastic neighbor embedding (t-SNE) [108] to visualize data samples and optimal transport dataset distance (OTDD) [5] to quantify dataset similarity.

t-SNE [108]. A widely used dimensionality reduction technique used to visualize high-dimensional data as a low-dimensional embedding. Initially, the algorithm converts data similarities to joint probabilities and minimizes KL-divergence between the low dimensional embedding and the high dimensional data [51].

OTDD [5]. Given two datasets with feature-label pairs, the distance between datasets is computed using theoretical underpinnings of optimal transport theory. The metric used to compute distances between features (e.g., euclidean distance) is combined with the Wasserstein distance between label distributions (over features). Thus, yielding a transportation 'cost' between the datasets, which is optimized as the lowest cost to couple data samples.

4.2 Traditional Machine Learning

In mental health research, traditional machine learning methods are often preferred because of their interpretability and applicability to smaller datasets [30, 104]. Moreover, these methods could be a good benchmark for deep learning approaches.

4.2.1 *Feature engineering & selection.*

The Audio/Visual Emotion and Depression Recognition Challenge (AVEC) provides a comprehensive list of speech-based features used to detect mental illness [107]. Many studies on speech-based depression [28, 71, 79, 88, 104, 110] and SI [104] use AVEC for analysis. We use the AVEC2013 feature set [107] to extract 2268 handcrafted features from the raw audio data using the openSMILE package [32]. Importantly, the Student study [104] released AVEC2013 feature set instead of raw audio to protect participant privacy, thus, we extract the same features to evaluate cross-dataset performance. It is vital to reduce the feature set size to ensure optimal training. After extracting the 2268 features, we used the mutual information (MI) metric [56] to reduce the number of features. MI computes a non-negative value that signifies the dependence between the feature and the discrete binary label [56]. Larger values indicate more dependence and thus could be more useful for prediction.

4.2.2 *Machine learning methods.*

In our analysis, we validated the performance of SI prediction using four machine learning approaches: (1) **Support vector machines (SVM)** [20]: a large-margin classifier capable of handling high dimensional data, (2) **Logistic regression (LR)** [48]: an effective statistical approach that assumes linearity, (3) **Random forest (RF)** [12]: a tree-based ensemble machine learning method that extends on decision trees (DT) through bagging, and (4) **Extreme gradient boosted trees (XGB)** [18]: a tree-based method that extends DT through boosting. For implementation, we first apply the mutual information metric to reduce the feature set. Next, the features are standardized. Finally, we perform a parameter search as described in Appendix A. Note that we describe train and test set splitting strategies in section 5.

4.3 Deep Learning

4.3.1 *Architectures.*

In contrast to feature engineering in traditional ML models, deep learning methods automatically generate feature embeddings for classification. Importantly, the embeddings are low-dimensional representations of the input audio signal, thus enabling us to compute similarity metrics or apply DA approaches efficiently. In deep learning for speech processing, the raw audio data is transformed into a log mel-spectrogram for training. The VGGish [45] architecture is a multi-layer convolutional neural network model trained on the YouTube 8M dataset [2] for large-scale audio classification. It takes log-mel spectrograms as input and generates embeddings $Z \in \mathbb{R}^{k\times128}$. In our analysis, we fine-tune (layers are frozen) the VGGish model resulting in the following variants:

(1) **VGGish-Z**. To leverage temporal dependencies from VGGish embeddings $Z \in \mathbb{R}^{k\times128}$, we fine-tune using an LSTM resulting in embeddings $L \in \mathbb{R}^{1\times128}$. Next, two fully-connected layers take L as an input for SI classification. We refer to this as VGGish-Z because the "intermediate input" to the LSTM is the VGGish embedding Z.

(2) **VGGish-L**. As the speech samples from VGGish-Z are variable length sequences, we extract LSTM embedding L from VGGish-Z for use in different DA approaches.

4.3.2 Implementation details. We implement deep learning models using pytorch, tensorflow, and keras. The models are trained for 500 epochs with a batch size of 32 using the categorical cross entropy loss function with the adam optimizer (lr=1×10^5). Moreover, to prevent overfitting we use earlystopping with a patience=25 and model checkpointing that restores the best model weights. Additional information is presented in the Appendix B.

4.4 Domain adaptation & domain generalization

To improve generalization capabilities, many DA methods have been proposed as discussed in section 2.3. While the specific implementation details are presented in Appendix B.2 & B.3, we briefly describe the UDA and SSDA methods used in our analysis:

(1) **Subspace Alignment (SA)** [33]: The method seeks to align the source and target domains by learning a mapping function between their respective subspaces. Thus, a transformation matrix M is learned to align source X_S and target X_T feature spaces. SA is a simple and effective method for domain adaptation, and using subspaces for out-of-distribution alignments has been explored in speech recognition [50, 63].

(2) **Linear Discrepancy Minimization (LDM)** [67]: It is an instance-based DA method where the emphasis is on data rather than features. Here, the source data is re-weighted by minimizing the linear discrepancy between the two domains.

(3) **Adversarial Discriminative Domain Adaptation (ADDA)** [105]: This adversarial framework is trained as follows. First, a source encoder generates good features for the specific task on the source domain. Next, a task network is trained using the source encoder to learn the task. Finally, a target encoder is trained to deceive a discriminator network that attempts to distinguish between source and target data in the encoded space.

(4) **Margin Disparity Discrepancy (MaDD)** [117]: Zhang et al. [117] introduced MaDD for unsupervised DA as a method with theoretical guarantees. Empirically, the technique is modified into an adversarial learning problem to learn a new feature representation that minimizes the discrepancy between source and target domains.

(5) **Attract, Perturb, and Explore (APE)** [55]: APE consists of three procedures. First, the target distribution discrepancy is minimized to globally align the target sub-distributions. Second, these distributions are further perturbed to accommodate unaligned target distributions. Third, the exploration procedure locally modulates the class-centers to enable more perturbation into the unaligned regions.

(6) **Contrastive Learning for DA (CLDA)** [95]: The CLDA framework proposes: (1) an instance contrastive alignment loss procedure between the unlabeled target samples and their augmented versions, and (2) an inter-domain contrastive alignment between the labeled source data and the prediction on unlabeled samples.

(7) **Entropy Minimization (ENT)** [41]: A network is trained to minimize the entropy on unlabeled target samples. Thus, clustering samples around a class center.

(8) **Minimax Entropy (MME)** [91]: This adversarial framework has two steps. First, the representative class sample (prototype) is updated by maximizing the entropy on the unlabeled target dataset. Second, the entropy is minimized to cluster features around the prototype, thus reducing distance between prototype and unlabeled samples.

4.5 Our Sampling Approach: Sinusoidal Similarity Sub-sampling (S3)

Motivation. We combine existing ideas in machine learning to address the unique challenges of mental health datasets. (1) *Small datasets*: The methods highlighted in section 4.4 are tailored for large datasets. Notably, mental health datasets presented in Table 1, have sizes less than 10^3, contrasting the 10^5 sizes of DA benchmark datasets like DomainNet [84]. S3 adopts a training-free metric solution to select best source samples for the target dataset, circumventing dataset size constraints. This idea stems from sample selection which is built-in in approaches such as OTDD [5], MME [91], and ENT [41]. (2) *Lack of labels & data-centricity*: The expense of data labeling impedes robust evaluation of models in mental health and healthcare. Data-centric methods have gained traction over training-based solutions because of their ability to generalize well using only features [118]. For example, methods such as Simi-Feat [118] rely on computing metrics using only the features, followed by a clustering approach to group similar samples. Following this, S3 computes scores between source and target pairs based solely on their features to detect similar source and target samples. Importantly, our experiments indicate that approaches with built-in sampling and metric computations are more effective for mental health (see sections 6.4 & 7).

S3 is based on the following ideas in speech and signal processing. First, Fourier methods assume that a signal can be decomposed into several sinusoidal signals [68]. The short-term fourier transform (STFT) computes sinusoidal frequencies of a signal as a function of time, yielding a spectrogram [68]. Second, spectrograms can be scaled based on the human perception of sound, obtaining log mel-spectrograms that capture time-frequency dynamics from speech samples [19, 98]. These are used as inputs in many large-scale audio classification models [45]. Hence, the embeddings generated by VGGish are latent spaces with time-frequency information. Using the above-mentioned principles, S3 computes a metric by comparing the source and target embedding to an *anchor* matrix Λ composed of randomly generated sine waves. Importantly, we construct Λ based on two factors. First, we assume the frequencies of the sine waves are between 80 to 250, covering the average range of human voice [6]. Second, the product of sine waves of different frequencies are orthogonal, which is analogous to vector orthogonality [92]. Consequently, the product of Λ with source and target embeddings captures frequency information present in both datasets. Now, we formalize our approach.

Problem Statement. Given datasets from the source domain $\mathcal{D}_S$ and the target domain $\mathcal{D}_T$ with sample-label pairs and samples $(\mathbf{x_s}, \mathbf{y_s})$ and $\mathbf{x_t}$, respectively, where $\mathbf{x} \in \mathbb{R}^k$ is an input audio signal of k seconds. S3 computes pair-wise scores $\gamma(\mathbf{x_s}, \mathbf{x_t})$ using the source and target embeddings of samples in $\mathcal{D}_S$ and $\mathcal{D}_T$.

As described in Algorithm 1, we compute score $\gamma(\mathbf{x_s}, \mathbf{x_t})$ in three stages:

(1) The embedding $Z \in \mathbb{R}^{n \times m}$ of sample $\mathbf{x}$ from a VGGish model is transformed into a sinusoidal matrix Φ using equation 2. Φ is composed of n sinusoidals of length m, where $\mathbf{k}_i^z$ is the i^{th} column vector of Z. Φ_{x_s} and Φ_{x_t} represent sinusoidal matrices for a source (x_s) and target (x_t) sample, respectively.

(2) To compute the *anchor* sinusoidal matrix Λ, we sample $\mathbf{k}^\lambda \in \mathbb{R}^{1 \times m}$ where each value $k \in [0, 2\pi]$. Next, Λ is generated using equation 3.

(3) The source (Φ_{x_s}) and target (Φ_{x_t}) sinusoidal matrices are multiplied with the anchor matrix (Λ), and the outputs are multiplied with each other and aggregated to compute the scalar score γ as shown in equation 4.

Algorithm 1 Computing S3 scores

```
Input: Source dataset $\mathcal{D}_S = \{(x_s, y_s)\}_{s=1}^{S}$, Target dataset $\mathcal{D}_T = \{x_t\}_{t=1}^{T}$, VGGish model with pre-processing $M$,
    vector of anchor frequencies $\mathbf{f}$.
Output: Pair-wise scores $\Gamma \in \mathbb{R}^{S \times T}$
    $M \leftarrow$ initialize weights
    for $s = 1$ to $S$ do
        for $t = 1$ to $T$ do
            $Z_s \leftarrow M(x_s)$
            $Z_t \leftarrow M(x_t)$
            for $i = 1$ to num_columns($Z_s$) do
                $\mathbf{k}_i^z \leftarrow Z_{s(i)}$
                $\Phi_{x_s} \leftarrow \mathbf{w}(\mathbf{f}, \mathbf{k}_i^z)$
            end for
            for $i = 1$ to num_columns($Z_t$) do
                $\mathbf{k}_i^z \leftarrow Z_{t(i)}$
                $\Phi_{x_t} \leftarrow \mathbf{w}(\mathbf{f}, \mathbf{k}_i^z)$
            end for
            Sample $\mathbf{k}^\lambda \in [0, 2\pi]$
            $\Lambda \leftarrow \mathbf{w}(\mathbf{f}, \mathbf{k}^\lambda)$
            $\gamma(\mathbf{x_s}, \mathbf{x_t}) \leftarrow \sum^m (\Lambda^T \Phi_{x_s}) \times (\Lambda^T \Phi_{x_t})$
        end for
    end for
```

$$\mathbf{w}(f, \mathbf{k}) = \sin(2\pi f \mathbf{k}) \tag{1}$$

$$\Phi(\mathbf{f}, \mathbf{k^z}) = \{\mathbf{w}(f_1, \mathbf{k}_1^z), \mathbf{w}(f_2, \mathbf{k}_2^z), \cdots, \mathbf{w}(f_n, \mathbf{k}_m^z)\} \tag{2}$$

$$\Lambda(\mathbf{f}, \mathbf{k}^\lambda) = \{\mathbf{w}(f_1, \mathbf{k}^\lambda), \mathbf{w}(f_2, \mathbf{k}^\lambda), \cdots, \mathbf{w}(f_n, \mathbf{k}^\lambda)\} \tag{3}$$

where $\mathbf{f} \in \mathbb{R}^{n \times 1}$ and each value $f \in [80, 250]$, $\mathbf{k}_i^z$ is the i^{th} column vector of Z, and $\mathbf{k}^\lambda \in \mathbb{R}^{1 \times m}$ where each value $k \in [0, 2\pi]$.

$$\gamma(\mathbf{x_s}, \mathbf{x_t}) = \sum^{m} (\Lambda^T \Phi_{x_s}) \times (\Lambda^T \Phi_{x_t}) \tag{4}$$

where $\Lambda^T \in \mathbb{R}^{m \times n}$ and $\Phi_{x_s}, \Phi_{x_t} \in \mathbb{R}^{n \times m}$

Empirically, we observed that S3 captures directional information in different ways (Fig. 6). Therefore, the best fine-tuning subset for the target domain is selected in three ways (Algorithm 2). For every target sample $\mathbf{x_t}$, **S3N** and **S3M** select optimal source sample using the smallest and largest $\gamma(\mathbf{x_s}, \mathbf{x_t})$, respectively. By relaxing the assumption that there is one-to-one correspondence between source and target sample, **S3R** selects two samples with smallest and largest pair-wise scores. In contrast to UDA approaches that emphasize feature transformations and sample re-weighting, S3 simply selects the optimal subset. Moreover, *S3 does not require labeled target data*, differing from other SSDA methods. Thus, making it suitable for smaller datasets in the mental health domain. We fine-tune the models using the best subset for 50 epochs using the same setup as deep learning models with earlystopping and model checkpointing.

Algorithm 2 Retrieve best samples using S3

```
Input: Source dataset $\mathcal{D}_S = \{(x_s, y_s)\}_{s=1}^{S}$, Target dataset $\mathcal{D}_T = \{x_t\}_{t=1}^{T}$, Pair-wise scores $\Gamma \in \mathbb{R}^{S \times T}$
Output: Best source dataset $\mathcal{D}_S^{best} \subset \mathcal{D}_S$
  for $t = 1$ to $T$ do
      for $s = 1$ to $S$ do $\mathcal{D}_S^{best} \leftarrow x_s$ such that $min/max/both(\Gamma(x_s, x_t))$
      end for
  end for
```

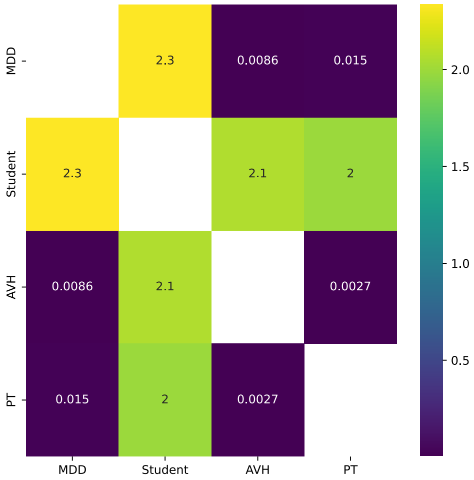


Fig. 2. Pair-wise dataset distances (divided by 10^{25}) computed using optimal transport dataset distance (OTDD). Smaller values indicate the datasets are more similar. AVH and PT datasets are similar to each other, whereas the Student dataset is dissimlar to all other datasets.

5 ANALYSIS

We divide our systemic analysis into four phases, each providing insights on generalization of speech-based detection. They are as follows:

(1) We perform dataset similarity evaluation using techniques from section 4.1 to understand quantitatively and qualitatively which data might generalize better (section 5.1).
(2) We compute baselines for SI detection by training and testing on the same dataset. Thus, helping us comprehend decreases and increases in subsequent analyses (section 5.2).
(3) We evaluate SI detection performance when models are trained on one dataset and tested on another. We refer to this as one-one validation. This helps us understand if specific populations or data subsets can be used to model the variance in other populations (section 5.3).
(4) We perform leave-one-dataset-out validation where models are tested on one dataset and trained on all other datasets. Consequently, this experiment answers the following questions: (1) Is more data better?, (2) Are deep learning models superior in many cases, and (3) How effective are UDA and SSDA approaches in handling dataset shifts?. In summary, it explores the trade-off between data quantity, data quality, and modeling (section 5.4).

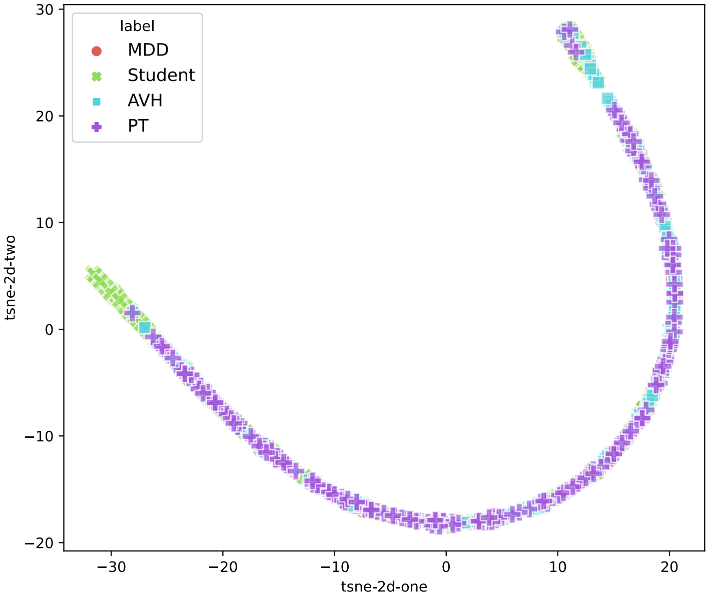


Fig. 3. t-SNE visualiztion of the four datasets: MDD, Student, AVH, and PT.

5.1 Characteristics of datasets

To evaluate data similarity, we compute the OTDD across the datasets using the 2268 AVEC2013 features extracted from raw audio data. An important advantage of this approach is its ability to quantify similarity. From Fig. 2, we observe that student is dissimilar to more clinical datasets such as MDD, AVH, and PT. Furthermore, AVH and PT have the smallest OTDD of 2.7×10^{28}. In fact, auditory verbal hallucinations and persecutory thoughts are common symptoms of psychotic disorders, including schizophrenia, bipolar disorder, and schizoaffective disorder [87, 106]. Capturing this "common-ness" using audio features motivates us to evaluate the predictive performance across these datasets. Furthermore, we observe that MDD is closer to AVH and PT. Recall that MDD is a clinical dataset, and its similarity to AVH and PT might be related to clinical sub-populations in these studies. It is worth noting that we refer to symptoms and not a diagnosis following the philosophy underlying the NIMH Research Domain Criteria (RDoC) [25, 34].

We use t-SNE [108] to visualize our dataset as a low-dimensional embedding. In addition to indicating the similarity of data samples, t-SNE plot also captures the variance in the dataset as a whole. Perplexity is an important t-SNE parameter that guesses the number of neighbors around a point. Thus, smaller and larger values emphasize local and global attention, respectively. As recommended in literature [108], we considered perplexity values between 5 and 100. After visually interpreting many figures, we chose perplexity=50 and iterations=2000. We refer to areas of t-SNE using (x, y) coordinates. From Fig. 3, we observe that the student dataset is more precise and less variant (-30, 5), suggesting that it might be difficult to generalize without diverse model representations. Moreover, larger datasets such as PT and AVH have more variance and span across the entire dataset spectrum, suggesting that they might be good candidates for training generalized models. As a smaller dataset, MDD will be a good candidate for testing; however, it might be difficult to train as the samples cannot adequately represent other larger datasets.

Table 3. Within-Dataset performance using balanced accuracy.

Model	Balanced Accuracy (std)			
	MDD	Student	AVH	PT
SVM	0.51 (0.02)	0.55 (0.05)	0.49 (0.01)	0.51 (0.02)
LR	**0.62 (0.02)**	**0.62 (0.08)**	0.50 (0.05)	0.56 (0.02)
RF	0.59 (0.20)	0.56 (0.05)	0.55 (0.04)	0.55 (0.07)
XGB	0.59 (0.20)	0.58 (0.09)	0.54 (0.05)	0.55 (0.04)
VGGish-Z	0.49 (0.02)	N/A	**0.62 (0.04)**	**0.68 (0.07)**
VGGish-Z + LR	0.40 (0.10)	N/A	0.59 (0.04)	0.52 (0.02)
VGGish-Z + RF	0.50 (0.00)	N/A	0.61 (0.03)	0.49 (0.04)

Table 4. Within-Dataset performance using F1-score and recall.

Model	F1-score (std)				Recall (std)			
	MDD	Student	AVH	PT	MDD	Student	AVH	PT
SVM	0.12 (0.10)	0.16 (0.15)	0.70 (0.08)	0.76 (0.15)	**0.60 (0.49)**	0.11 (0.11)	**0.86 (0.19)**	0.82 (0.27)
LR	**0.23 (0.29)**	**0.35 (0.10)**	0.57 (0.02)	0.69 (0.04)	0.40 (0.49)	**0.53 (0.20)**	0.54 (0.02)	0.61 (0.08)
RF	0.20 (0.40)	0.35 (0.15)	0.69 (0.05)	0.82 (0.04)	0.20 (0.40)	0.27 (0.16)	0.77 (0.07)	0.90 (0.05)
XGB	0.20 (0.40)	0.30 (0.21)	0.68 (0.04)	0.83 (0.03)	0.20 (0.40)	0.26 (0.20)	0.73 (0.04)	0.91 (0.04)
VGGish-Z	0.00 (0.00)	N/A	0.49 (0.04)	0.77 (0.07)	0.00 (0.00)	N/A	0.35 (0.05)	0.71 (0.04)
VGGish-Z + LR	0.00 (0.00)	N/A	0.65 (0.60)	0.70 (0.04)	0.00 (0.00)	N/A	0.62 (0.11)	0.63 (0.07)
VGGish-Z + RF	0.00 (0.00)	N/A	**0.71 (0.04)**	**0.83 (0.02)**	0.00 (0.00)	N/A	0.75 (0.10)	**0.93 (0.02)**

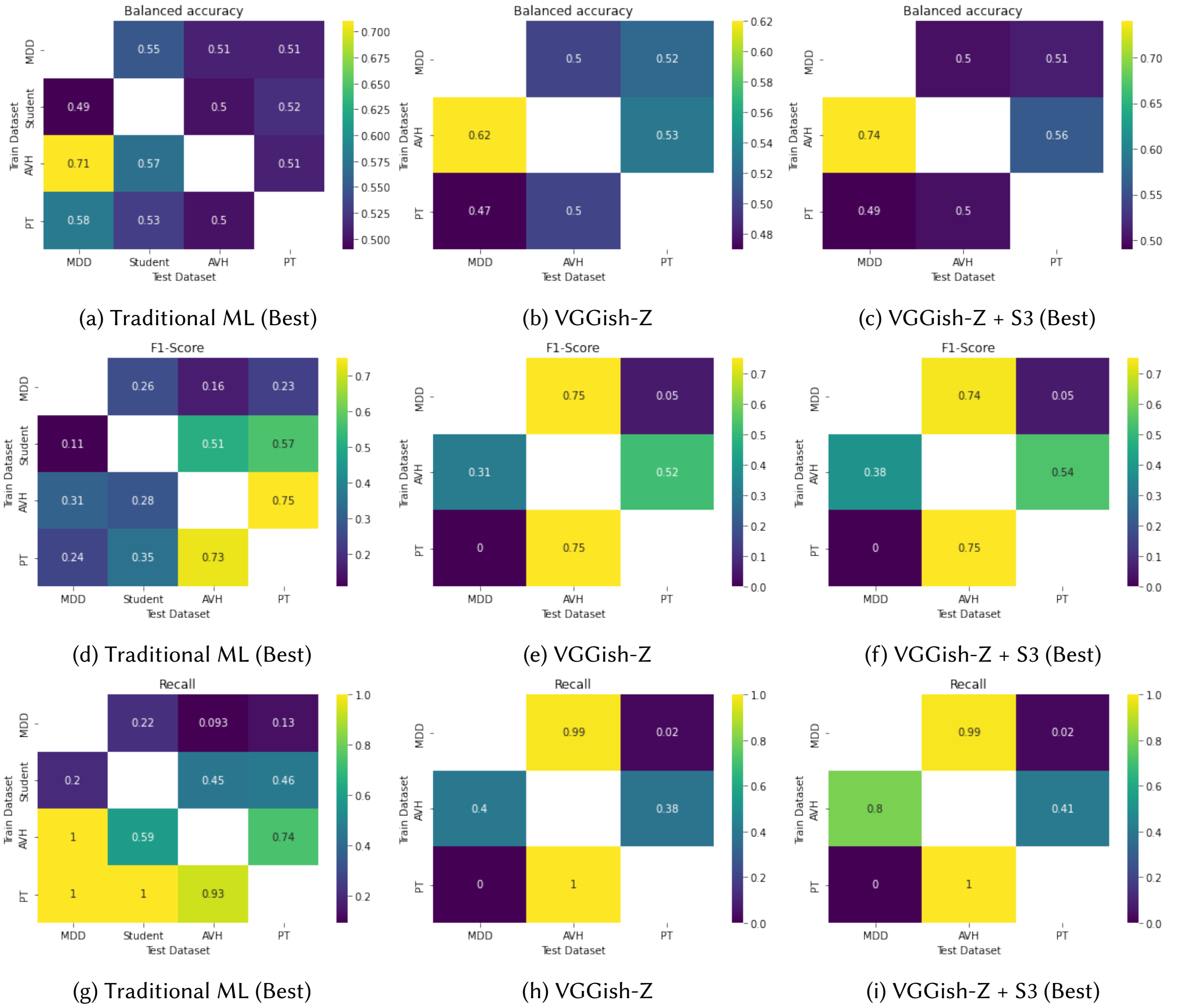
5.2 Evaluating within-dataset performance

Establishing within-dataset performance baselines is a crucial prerequisite for evaluating generalization. Here, we train and test the models on the same dataset using stratified-5-fold cross-validation. From table 3, we observe logistic regression obtained a balanced accuracy of 0.62 for the MDD and Student datasets and poor results for AVH and PT. In contrast, VGGish-Z obtained a balanced accuracy of 0.62 and 0.68 for AVH and PT, respectively, and performed poorly for other datasets. However, we ask, "Are larger and balanced datasets better?". Recall that AVH and PT have 356 (SI=60.1%) and 209 (SI=74.6%) participants, respectively. From Table 3, we observe that AVH has lower balanced accuracy than PT (0.62 vs 0.68), suggesting that factors other than data size and balance are important. Perhaps, AVH has a more heterogeneous cohort than PT, and capturing their characteristics is harder. In essence, data choice is a crucial component.

Furthermore, we use the F1-score and recall are used to evaluate performance on positive SI detection. By comparing tables 3 and 4, we observe that the models with the highest F1 scores align consistently with balanced accuracy. Specifically, LR achieves the best scores for MDD (0.23) and Student (0.35), while the VGGish-Z-based model achieves the highest score for AVH (0.71) and PT (0.83). However, we notice a trade-off between balanced accuracy and recall. In particular, SVM attains the best recall scores for MDD (0.60) and AVH (0.86) but shows poor balanced accuracy of 0.51 and 0.49, respectively.

In summary, we observe that relatively small and homogeneous datasets such as MDD (clinical) and Student (non-clinical) can be modeled better using traditional ML models. In contrast, deep learning methods perform better on large heterogeneous datasets such as AVH (mixed) and PT (mixed). As the student dataset's raw audio is not released to protect personal identifiable information, we do not test it with deep learning methods.

5.3 One-One Validation



(a) Traditional ML (Best)

(b) VGGish-Z

(c) VGGish-Z + S3 (Best)

(d) Traditional ML (Best)

(e) VGGish-Z

(f) VGGish-Z + S3 (Best)

(g) Traditional ML (Best)

(h) VGGish-Z

(i) VGGish-Z + S3 (Best)

Fig. 4. One-One evaluation with balanced accuracy (top row), F1 (middle row), and recall (bottom row). Best refers to the top performing method. The scores for each model/method are shown in Appendix C.

As a first step toward evaluating generalizability, we train our models on one dataset and test it on another. We refer to this setup as one-one validation and represent our results as a matrix. From Fig. 4, we make many interesting observations as follows.

First, out-of-distribution performance is lower than within-dataset in many cases, as shown in Fig. 4 (a), (b), (d), (e). Student, AVH, and PT have lower balanced accuracy scores with decreases $\Delta = -0.05$, $\Delta = -0.12$, and

Table 5. Leave-one-dataset-out validation using traditional learning methods.

Framework	Model	Balanced Accuracy				F1-Score (Recall)			
		MDD	**Student**	**AVH**	**PT**	**MDD**	**Student**	**AVH**	**PT**
Base	Benchmark (Table 3, 4)	0.62	0.62	0.55	0.56	0.23 (0.60)	0.35 (0.53)	0.70 (0.86)	0.83 (0.91)
	Benchmark (Figure 4)	0.71	0.57	0.51	0.52	0.31 (1.00)	0.35 (1.00)	0.73 (0.93)	0.75 (0.74)
	SVM	0.56	0.50	0.49	0.50	0.23 (0.40)	0.34 (0.81)	0.75 (0.99)	0.00 (0.00)
	LR	0.62	0.50	0.51	0.46	0.25 (0.60)	0.30 (0.59)	0.74 (0.99)	0.65 (0.45)
	RF	0.52	0.54	0.50	0.46	0.21 (0.40)	0.35 (0.81)	0.68 (0.89)	0.67 (0.47)
	XGB	0.56	**0.56**	**0.55**	0.49	0.23 (0.40)	0.36 (1.00)	0.69 (0.89)	0.20 (0.13)
UDA	LR + LDM	0.46	0.54	0.50	**0.52**	0.16 (0.40)	0.30 (0.62)	0.75 (0.99)	**0.71 (0.84)**
	RF + LDM	**0.64**	0.54	0.50	0.47	**0.27 (0.60)**	0.36 (1.00)	0.73 (0.99)	0.65 (0.47)
	LR + SA	0.60	0.54	0.46	0.46	0.25 (0.60)	0.30 (0.59)	0.47 (0.45)	0.46 (0.38)
	RF + SA	0.56	0.54	0.50	0.50	0.23 (0.40)	0.36 (1.00)	0.00 (0.00)	0.00 (0.00)

$\Delta = -0.15$, respectively. Moreover, notice that deep models have more severe reductions than ML methods. Perhaps, the representations learned by deep methods are tuned for specific cohorts. From Fig. 4, we notice that models trained on MDD do not transfer well to other datasets. The MDD dataset size and balance is the most probable reason for this result. Moreover, we observe that large diverse datasets such as AVH and PT generalize better without any adaptation or tuning. In particular, AVH is the best dataset for generalization with balanced accuracies of 0.71, 0.57, and 0.53 for MDD, Student, and PT, respectively.

Second, from Fig. 4 (d), (e), (g), (h), we observe that PT exhibits higher positive predictive power compared to other datasets. Utilizing traditional ML methods, PT achieves recall scores of 1, 1, and 0.93 for MDD, Student, and AVH, respectively. However, its precision is relatively low, as indicated by the corresponding F1-scores of 0.24, 0.35, and 0.73. Additionally, AVH generalizes well to PT, whereas the opposite is not true, as evidenced by the balanced accuracies of the three models. In summary, our findings underscore the significance of identifying samples that can generalize to the target data, which serves as a motivation for the development of S3.

Finally, S3 improves the generalization of deep learning models in many cases. In particular, AVH to MDD, AVH to PT, PT to MDD (Fig. 4 (b), (c), (e), and (f)). Fine-tuning AVH for MDD using S3 variants improves balanced accuracy by $\Delta = 0.12$ and F1 by $\Delta = 0.07$ over a standard VGGish-Z deep model. In fact, these scores are better than within-dataset performance for MDD (balanced accuracy: $\Delta = 0.12$; F1: $\Delta = 0.17$). These results suggest that S3 is well-suited to transfer performance from larger to smaller datasets in the most optimal way. Moreover, notice that S3 yields performance improvements when AVH is fine-tuned for PT (balanced accuracy: $\Delta = 0.03$; F1: $\Delta = 0.02$). In contrast to methods that emphasize model tuning, S3 improves performance by choosing optimal samples for fine-tuning, thus, highlighting the strengths of data-centric approaches.

5.4 Leave-one-dataset-out validation

In this experiment, we evaluate generalization in leave-one-dataset-out (LODO) validation, where one dataset is used for testing, and all others are used for training. Here, we evaluate UDA and SSDA to improve generalization.

From Table 5, we observe that, in five out of the eight cases, LODO further decreases the balanced accuracy of traditional ML compared to one-one validation benchmark (Figure 4). In particular, balanced accuracy in MDD ($\Delta = -0.09$), Student ($\Delta = -0.01$) and PT ($\Delta = -0.02$), and F1 scores in MDD ($\Delta = -0.06$) and PT($\Delta = -0.08$). Therefore, we fine-tuned the ML methods using linear discrepancy minimization (LDM) and subspace alignment (SA) to accommodate distribution shifts in the target domain. DA methods improved performance over traditional ML in 4 cases: balanced accuracy of MDD ($\Delta = 0.02$) and PT ($\Delta = 0.02$), and F1-scores of MDD ($\Delta = 0.02$) and PT

Table 6. Leave-one-dataset-out validation using deep learning, UDA, and SSDA approaches.

Framework	Model	Balanced Accuracy			F1-Score (Recall)		
		MDD	AVH	PT	MDD	AVH	PT
Base	Benchmark (Table 3, 4)	0.49	0.62	0.68	0.00 (0.00)	0.71 (0.75)	0.83 (0.93)
	Benchmark (Fig. 4)	0.62	0.50	0.53	0.31 (0.40)	0.75 (1.00)	0.52 (0.38)
	VGGish-Z	0.65	0.51	0.49	0.31 (0.60)	0.71 (0.84)	0.85 (0.99)
UDA	MaDD [117]	0.50	0.50	0.50	0.21 (1.00)	0.75 (1.00)	0.85 (1.00)
	ADDA [105]	0.50	0.50	0.50	0.21 (1.00)	0.75 (1.00)	0.85 (1.00)
SSDA	Random	0.63	0.50	0.49	0.28 (0.40)	0.75 (0.95)	0.84 (0.95)
	APE [55]	0.50	0.50	0.50	0.26 (1.00)	**0.77 (0.98)**	0.85 (1.00)
	CLDA [95]	0.50	0.51	0.50	0.26 (1.00)	0.75 (0.98)	0.85 (1.00)
	ENT [41]	0.63	0.50	0.50	0.33 (0.80)	0.75 (1.00)	0.85 (1.00)
	MME [91]	0.59	0.51	**0.53**	0.31 (0.40)	0.76 (0.96)	0.78 (0.85)
	S3N (proposed)	0.67	0.52	0.50	0.33 (0.60)	0.75 (0.98)	**0.86 (1.00)**
	S3M (proposed)	0.66	0.51	0.50	0.32 (0.60)	0.74 (0.95)	0.84 (0.95)
	S3R (proposed)	**0.68**	**0.52**	0.52	**0.35 (0.60)**	0.76 (0.99)	0.83 (0.92)

(Δ = 0.04). Moreover, logistic regression with LDM on AVH performs better on the SI class than the benchmark indicated by the 0.75 F1-score (Δ = 0.02). Surprisingly, a random forest with LDM obtained a balanced accuracy of 0.64, surpassing the within-dataset benchmark (Δ = 0.02). Logistic regression with SA and LDM performs reasonably well for adaptation. Perhaps, the combination of using a linear model, linear adaptation, and a small dataset is well-suited for SA. To summarize, our results suggest that domain adaptation works in 50% of the cases. Importantly, using AVH, a large, diverse dataset, as the source domain improves performance. Thus, indicating the importance of data choices.

In the base deep learning setup (Table 6), we observe that the smaller dataset (MDD:Δ = 0.16) benefits greatly from more diverse data, whereas larger datasets suffer in the LODO setup. Generalization performance decreases in two out of six cases: balanced accuracy of AVH (Δ = −0.11) and PT (Δ = −0.19). Consequently, we investigated UDA and SSDA method to improve cross-dataset performance. From Table 6, we observe that adversarial UDA such as MaDD and ADDA do not improve generalization, mainly because adversarial training requires large datasets [3].

Among SSDA approaches, we notice that APE and CLDA are less effective than ENT and MME. APE obtained the best F1 score (0.77) on AVH, nevertheless, it does not perform well on other datasets. Similarly, CLDA generalizes poorly to other SI datasets. Perhaps the training procedures for contrastive learning and APE are not viable in the context of SI detection. From Table 6, observe that ENT obtained the best balanced accuracy for MDD (0.63) compared to other SSDA baselines. Importantly, MME works across datasets with balanced accuracies of 0.59, 0.51, and 0.53 for MDD, AVH, and PT datasets.

From Table 6, we observe that S3 variants outperforms baselines across the datasets. In particular, in many cases, S3R improves performance over VGGish-Z with balanced accuracy (MDD: Δ = 0.03; AVH: Δ = 0.01; PT:Δ = 0.03) and F1 (MDD: Δ = 0.04; AVH: Δ = 0.05). S3R also outperforms the one-one benchmark in five scenarios concerning balanced accuracy (MDD: Δ = 0.06; AVH: Δ = 0.02) and F1 (MDD: Δ = 0.04; AVH: Δ = 0.01;

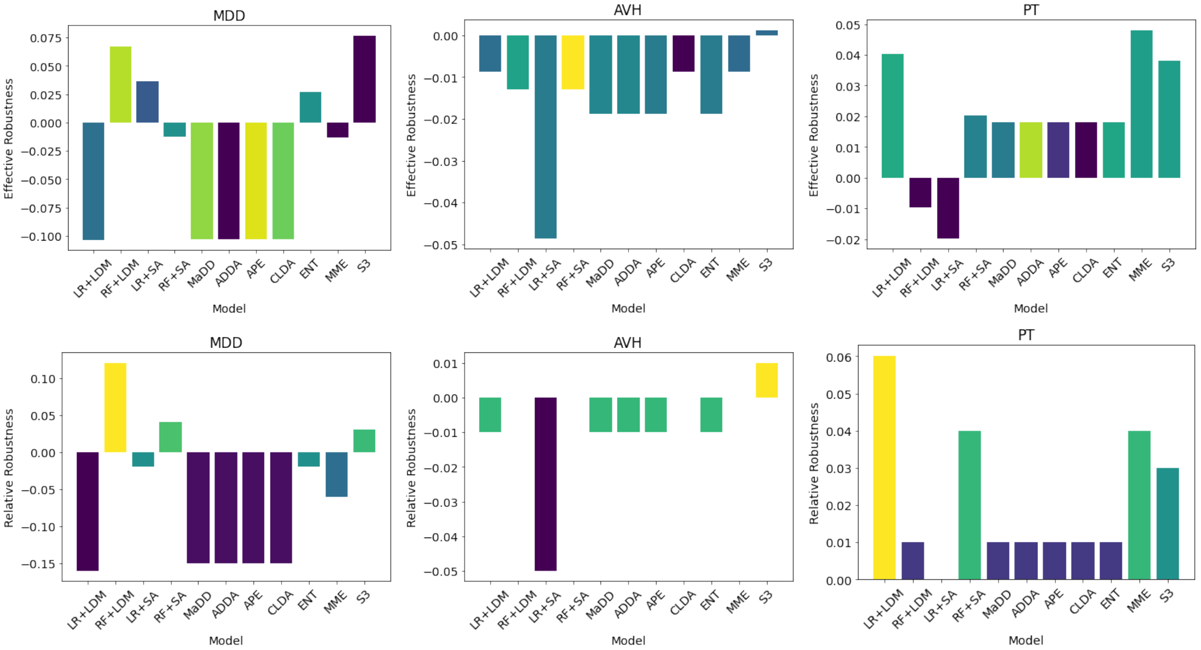


Fig. 5. Measuring effective robustness (top row) and relative robustness (bottom row) of UDA and SSDA baselines, and S3. Positive values indicate useful robustness to distribution shifts.

PT: $\Delta = 0.31$). S3 considerably outperforms SSDA approaches for the smallest dataset (MDD with N=43) obtaining a balanced accuracy of 0.68, while ENT obtained 0.63. In other cases with AVH and PT, S3 improves generalization incrementally. It is noteworthy that MME obtained the best generalization to PT (balanced accuracy = 0.53). We discuss these results from a mental health context in section 7.

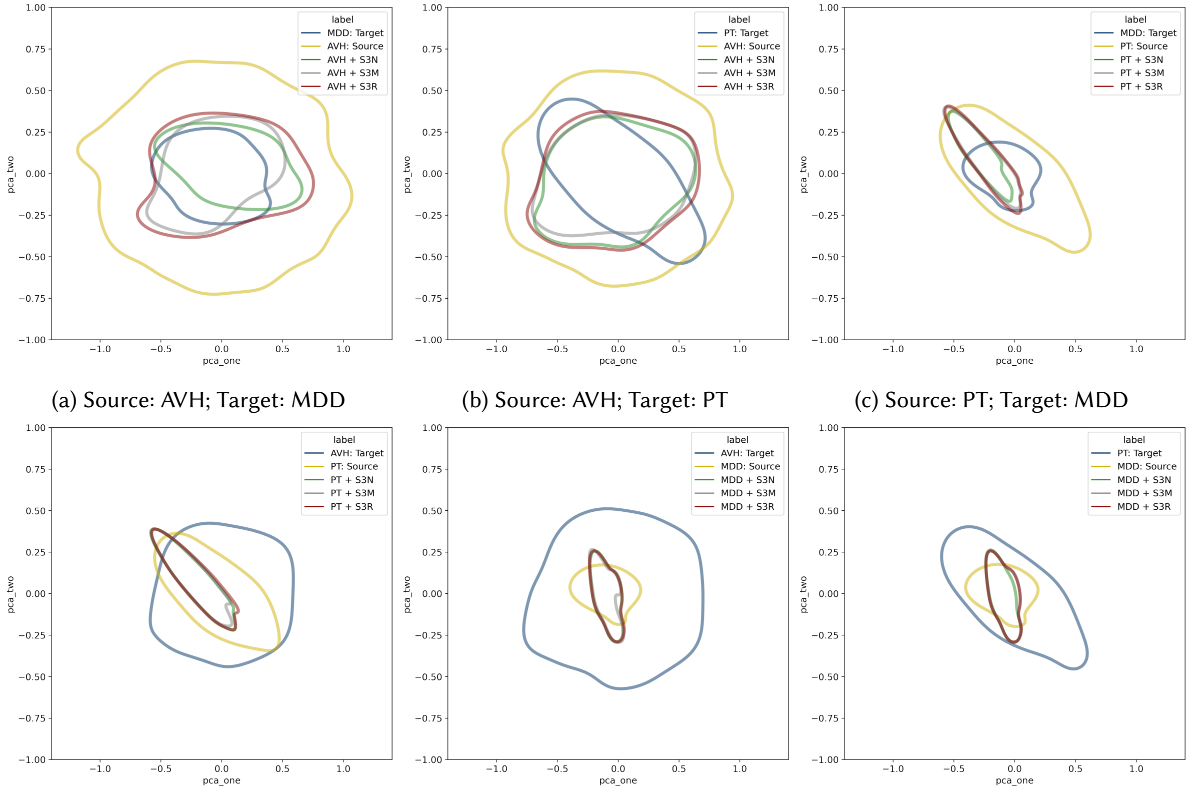
6 POST-HOC ANALYSIS

6.1 Result highlights by measuring robustness

We compile our analysis by measuring robustness, thus summarizing the performance of DA approaches across MDD, AVH, and PT. We disentangle accuracy from robustness using the notions of effective and relative robustness proposed by Taori et al. [102]. Effective robustness (ρ) quantifies if the accuracy under distribution shift is better than what is expected from obtaining higher within-dataset accuracy. Given a model M, we compute ρ using equation 5 as follows. First, β is computed using a log-linear on the base models without DA. Next, the slope and intercept are used to predict the expected values. Finally, the difference between these values and accuracy with DA is computed. Relative robustness (τ) is computed as the accuracy difference between the base model and DA in the LODO setting.

$$\rho = acc_{lodo}(M) - \beta(acc_{within}(M)) \tag{5}$$

LDM with traditional ML emerged as a robust UDA approach for the MDD dataset with a $\rho = 0.06$ & $\tau = 0.12$ (Fig. 5). However, deep UDA approaches had poor generalization for the MDD and AVH datasets. From Fig. 5, we can also observe that RF is particularly robust for MDD and PT exhibiting positive ρ and τ. Moreover, we observe



(a) Source: AVH; Target: MDD

(b) Source: AVH; Target: PT

(c) Source: PT; Target: MDD

(d) Source: PT; Target: AVH

(e) Source: MDD; Target: AVH

(f) Source: MDD; Target: PT

Fig. 6. Kernel density estimate plots describing the distribution of the data chosen by S3N, S3M, and S3R. Source and target domains are represented by the gold and blue contours, respectively. A compact fit to the target domain is more desirable. For example, in subfigure (a), we observe that S3R is compact and encompasses MDD better than S3N and S3M.

that *S3 obtains positive robustness values across all datasets* compared to other SSDA methods. In fact, it is the only method to achieve this in the AVH dataset. MME emerged as the most robust for PT ($\rho = 0.048$). Interestingly, most methods are not robust for AVH, whereas most methods are robust for PT. The most probable reason for this is that the effect of distribution shift is more drastic in PT than AVH. Thus, we expect more improvements in PT than AVH when applying a DA approach.

6.2 Closer examination of S3

A deeper investigation into the factors that contribute to S3's effectiveness is necessary. As S3 samples data for fine-tuning, it is natural to question if performance gains are from additional fine-tuning rather than chosen samples. Thus, we train with a random subset referred to as Random. From Table 6, we notice Random's performance deteriorates or remains the same in most cases (5 out 6 cases), suggesting that samples selected by S3 are conducive to fine-tuning models to the target domain. Next, we compare the probability distributions of the source and target datasets with subsets selected by S3N, S3M, and S3R. Here, we first apply PCA to the VGGish-L embeddings to obtain a 2D latent space. Next, we use kernel density estimation to model the probability distribution of the

Table 7. Top-3 most correlated handcrafted features with learned features of VGGish + S3R. The significance of learned features with labels is computed using Kruskal-Wallis Test at $p < 0.05$. Spearman ρ is used to compute correlations between significant and handcrafted features, and it is presented in braces.

MDD	**AVH**	**PT**
(+0.59) Mean peak distance Spectral roll-off @ 90%	(+0.35) Flatness Spectral roll-off @ 90%	(+0.46) Kurtosis F_0
(+0.58) Std peak distance Spectral roll-off @ 90%	(+0.35) Flatness Spectral roll-off @ 75%	(+0.44) IQR 1-2 frequency band 250-650Hz
(+0.58) 1st quartile Spectral roll-off @ 75%	(+0.33) Flatness psycho-acoustic sharpness	(+0.44) IQR 1-3 frequency band 250-650Hz
(-0.64) Rise time Spectral roll-off @ 50%	(-0.35) 3rd quartile MFCC_2	(-0.44) 1st quartile frequency band 250-650Hz
(-0.60) Skewness of MFCC_12	(-0.34) 99th percentile MFCC_2	(-0.42) 1st quartile Spectral harmonicity
(-0.60) 3rd quartile Spectral roll-off @ 50%	(-0.33) 99th percentile MFCC_4	(-0.41) IQR 1-3 Shimmer

different datasets. The contour plots of the 2D PCA are shown in Fig. 6. Here, the source and target domains are represented in gold and blue, respectively. Contours that obtain a compact fit on the target domain are desirable as they minimize variance unnecessary for the target domain.

From Fig. 6, we observe that S3 attempts to "pull apart" or "compress" the source distribution to fit the target domain. Notably, in Fig. 6 (a), the subset selected by S3 variants reduces the distribution of AVH to the model MDD domain. Moreover, we observe that S3N and S3M ignore some target samples and only fit MDD in one direction. However, S3R almost completely overlaps MDD considering both latent directions. We observe similar trends to varying degrees in all other cases. We notice that S3 performs well in scenarios where the source domain is larger (Fig. 6 top row), improving the compactness of the chosen subset. However, it cannot stretch the distribution boundary to generalize from smaller to larger datasets. This observation is intuitive as S3 only selects a subset of samples rather than transform source features. Nevertheless, we can still observe S3R stretches better than S3M and S3N.

6.3 Interpretability

Understanding the learned features of the deep learning model is vital for SI detection applications. In particular, we want to interpret features that are important for generalizability, thus we use models from the leave-one-dataset-out setting. To identify associations between the latent feature of VGGish+S3R (sections 4.3 & 4.5) and traditional handcrafted features (section 4.2.1), we use rigorous statistical tests in two stages:

(1) As interpreting variable length sequences is non-trivial, we extract the output of our penultimate fully connected layer to obtain a 128-D vector for each input. Next, we apply the Kruskal-Wallis Test (non-parametric version of ANOVA) to identify significant learned features between the SI and non-SI groups.
(2) We select the top three significant features and compute their spearman correlations with the AVEC2013 handcrafted features described in section 4.2.1.

We identified 3, 4, 9 significant features ($p < 0.05$) for the MDD, AVH, and PT datasets, respectively. The 3 most significant features in each dataset exhibiting the most positive and negative correlations with the handcrafted features are presented in Table 7. Here, we observe that spectral roll-off is crucial for detection across both MDD and AVH datasets. Interestingly, most spectral roll-off functionals are positively correlated whereas mel-frequency cepstrum coefficients (MFCC) are negatively correlated. Notice that the psycho-acoustic sharpness variable in

Table 8. Investigating domain adaptation techniques for acoustic scene classification. Within-Dataset refers to training and testing on the same dataset, whereas all other methods use LODO validation. The three domains A, B, C refer to different recording devices. Metrics are presented as the unweighted mean of the individual class metrics. Recall and precision are reported in Appendix E.

Framework	**Method**	**Full Dataset**						**10% of Dataset**					
		Accuracy			**F1-Score**			**Accuracy**			**F1-Score**		
		A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Within-Dataset	VGGish	0.68	0.49	0.40	0.67	0.47	0.39	0.51	0.27	0.27	0.49	0.17	0.20
LODO	VGGish	0.39	0.24	0.39	0.36	0.23	0.37	0.27	0.31	0.34	0.31	0.30	0.31
	APE [55]	0.18	0.17	0.14	0.06	0.06	0.05	0.17	0.20	0.20	0.06	0.06	0.09
	CLDA [95]	**0.36**	**0.39**	**0.39**	0.32	**0.37**	**0.36**	0.21	0.30	0.25	0.07	0.25	0.10
	ENT [41]	0.26	0.23	0.24	0.19	0.16	0.17	0.26	0.30	0.30	0.15	0.20	0.23
	MME [91]	0.35	0.25	0.26	**0.34**	0.18	0.19	**0.29**	0.30	0.32	0.25	0.15	0.29
	S3R (proposed)	0.35	0.24	0.36	0.33	0.23	0.33	0.27	**0.34**	**0.35**	**0.31**	**0.34**	**0.34**

AVH is correlated to our deep learning embeddings, suggesting that our models can capture acoustic measures of mental health symptoms. Important features for detection in PT are different from MDD and AVH, suggesting that some acoustic differences are present across different populations.

6.4 S3: Mental health vs. Other domains

Recall that S3 is designed to address mental health tasks which are characterized by small datasets and unlabeled target domain, benefiting data-centricity. On the opposite end, we wanted to examine if S3 can extend to large-scale audio datasets. Thus, we build models to perform acoustic scene classification (ASC) using the TAU urban acoustic scenes 2019 mobile development dataset [69]. Given a 10s audio sample, the goal of ASC is to classify it into one of 10 classes (airport, indoor shopping mall, metro station, pedestrian street, public square, street, tram, bus, underground metro, urban park). We tackle sub-task b which consists of three sub-datasets A , B, and C referring to data obtained from different recording devices/domains. Datasets A, B, and C consists of 14400, 1080, and 1080 samples, respectively. Importantly, to mirror mental health dataset sizes, we perform the same analysis with 10% of the dataset, i.e., A, B, and C consist of 1440, 108, and 108 samples, respectively. Evaluating the performance of SSDA approaches on the full and 10% datasets will highlight the advantages and disadvantages of S3 (Table 8). For brevity, training details are described inAppendix E.

Table 8 shows many interesting observations relevant to understanding S3. While S3 performs reasonable well for the full dataset, CLDA is clearly the best performing method. In contrast, S3 is the best performing method on the smaller dataset (10% of dataset). Other methods like MME and ENT also demonstrate notable effectiveness on this reduced dataset, an observation noted in SI detection. In summary, S3 is not an optimal choice for broad, large-scale SSDA applications, but it proves to be effective in the mental health domain, which typically involves smaller datasets.

7 DISCUSSION

7.1 Summary of results

In our within-dataset analysis (section 5.2), we find that deep learning and machine learning perform better with larger and smaller datasets, respectively. This finding has been observed by other studies examining the impact of dataset size on deep learning model performance [8, 101]. Moreover, in the StudentSADD study [104], the best performing traditional ML and deep model obtained balanced accuracy of 0.57 and 0.73, respectively. Similarly, we obtain balanced accuracy ranging from 0.62-0.68 for the different datasets, indicating the challenging nature of speech-based SI detection.

Using OTDD and t-SNE, we observed that AVH and PT are similar datasets, while Student is dissimilar to all other datasets. Generally, we assume similar datasets to transfer better. Through our one-one validation, we make several interesting observations. First, our assumption about data similarity transferring better is untrue. While AVH had minor generalization on PT, the inverse is not true. Second, transfer from larger to smaller datasets is better, as observed by other studies [116]. The observations above emphasize the need to choose the "right" data for adaptation, serving as a motivator for S3. Third, S3 variants performed the best, improving over VGGish-Z. In fact, using S3M with AVH for MDD obtained a balanced accuracy of 0.74, which is $\Delta = 0.25$ higher than the within-dataset baseline.

Through leave-one-dataset-out validation (section 5.4), we evaluate the performance of testing on one dataset while training on all other datasets. We observe that distribution shift leads to decreased performance in many cases in both traditional machine learning and deep learning. Many studies have studied the effects of distribution shifts of performance and propose DA and DG methods for mitigating their effects [39, 78]. We employ some of these methods to investigate their effectiveness in alleviating this problem. Subsequently, we observed that using LDM with machine learning methods mitigated performance decreases in many cases. For deep learning methods, we noticed that adversarial UDA approaches such as ADDA are insufficient to improve performance. In summary, UDA approaches work reasonably well for traditional ML approaches compared to deep models.

Among SSDA methods, S3 improves cross-dataset performance in most cases over the baselines. Our analysis attributes these improvements to the specific design elements of S3, which are useful for mental health applications. Recall that S3 leverages a metric-based solution to subsample from the source dataset. While ENT and MME demonstrate improved cross-dataset performance, CLDA and APE are ineffective. CLDA's contrastive learning framework relying on strong augmentations and abundant unlabeled training data, is not suited to address our domain's limitations. Similarly, APE's demand for matrix uniformity to assess maximum mean discrepancy poses challenges. ENT's core principle is to use conditional entropy [41] to discern samples beneficial for domain adaptation. This mirrors S3's metric-driven optimal sample selection. MME's approach hinges on measuring sample "distance" from class prototypes. From a feature perspective, this is similar to computing the distance between anchor and source/target embedding in S3 (equation 4). By comparing the presence/absence of components of SSDA baselines with S3 suggests that S3's design elements like subsampling, preference for metric computations over feature transformations, and data-centric focus are paramount when dealing with mental health datasets. Importantly, our empirical analysis on acoustic scene classification (section 6.4), indicates that S3 excels with smaller datasets. This makes it particularly suitable for mental health contexts, where small and often unlabeled datasets are common, and data-centric solutions are crucial.

We analyze the choice of the S3 variant by summarizing our analyses. Notice that S3M selects one optimal sample whereas S3R selects two, thus relaxing the assumption of only one optimal source sample. We observed that S3 variants improve generalization for one-one (section 5.3) and leave-one-dataset-out validation (section 5.4). Moreover, from Fig. 6, we see that S3M captures variance in one direction, while S3R accommodates larger subsets of data, suggesting that S3M and S3R are suited for single-domain and two-domain scenarios, respectively. Further investigation is necessary to evaluate S3R's effectiveness in multi-domain setups using many datasets. In our interpretability analysis, we find that spectral roll-off is important feature that generalizes across MDD and AVH. Previous studies have suggested the effectiveness of spectral roll-off for detecting depression [66, 99] and somatization disorder [86]. However, this feature was not important for PT, suggesting that differences exist between populations.

7.2 Implications

7.2.1 Human computer interaction. Collecting speech-based diaries from mobile phones is a crucial component of our work. Smartphones offer a fast and cost-effective way of administering interventions or monitoring at-risk individuals. Our investigation implicitly suggests that diverse smartphones instrumented with different microphones can feasibly detect suicidal ideation. Therefore, we can extend our study to provide just-in-time-adaptive interventions (JITAI) in two ways. First, we can integrate our method into existing applications such as Talkspace to screen individuals experiencing a mental health crisis and connect them to mental health experts. Second, we can provide personalized screening, integration, and self-monitoring to individuals by tracking their mental health history. During monitoring, mobile phones can deliver longitudinal interventions such as cognitive behavioral therapy (CBT) or mindfulness-based stress reduction (MBSR).

7.2.2 Resource-constrained populations. The under-detection of mental illness in resource-constrained environments is a common problem [52]. Many low and middle-income countries have a large psychiatric disease burden without sufficient resources [82]. Furthermore, many populations with SI are understudied. In such scenarios, data collection efforts lead to small imbalanced datasets, making computational modeling more challenging. Moreover, generating large labeled datasets is infeasible owing to large resource burdens. Our investigation directly addresses many of these challenges. For example, smaller target domains significantly benefit from optimal transfer using larger source datasets. These findings are important in many cases. For example, SI in understudied mental disorders such as body dysmorphia [85] can be analyzed with reduced data collection efforts. Similarly, extending our method based on socio-economic status and geography can help underrepresented groups. Following NIMH RDoC [25, 34], we evaluate a common symptom across populations, with the potential to advance understanding of these symptoms at a level that is more general than that of typically-used diagnostic categories.

7.2.3 Data-centric machine learning. Current machine learning research is model-centric, where improving predictive performance involves experimenting with new architectures, loss functions, optimization methods, etc. In contrast, data-centric machine learning refers to systemically engineering data in different ways to improve predictive performance. Here, the data is given more importance, and the models are assumed to be fixed. Recent work in this area focuses on covariate shifts and trustworthy data samples [17, 93]. We believe that robust in-the-wild systems in speech-based SI detection via mobile phones could benefit greatly from data-centric approaches. S3 is data-centric as it selects samples from the source domain that is more likely to explain shifts in the target domain. Furthermore, we evaluate performance on unseen users, where the model has no prior information about the user or the domain in many cases. Thus, aiding translation to real-world applications and addressing the cold-start problem [65].

7.2.4 Challenges of generalizability in mental health. Through comprehensive analysis of DA methods using multiple datasets, we achieve incremental improvement over many previous methods. Nonetheless, our results highlight that detecting rare symptoms in small mental health datasets with rare samples is challenging with great room for improvement. Here, we suggest some future directions to directly tackle this unresolved issue. First, as data is limited, models may benefit from incorporating expert knowledge instead of a purely learning based approach. Second, multi-modal methods could enhance the amount of data available. Moreover, using different modalities can complement each other and thus improve generalization. Third, we could personalize models to investigate if it works across people before we generalize across populations.

7.3 Limitations

Here we discuss the limitations of our study and proposed S3 algorithm. First, while we investigate generalizability across four datasets, it is crucial to understand the trade-offs with generalization. A method that works in all cases is impossible and might not be necessary. In our studies, we make many efforts to mitigate recruitment biases and include people representing the studied population. Nevertheless, biases could arise from many factors, including gender, race, geography, and socioeconomic status. Second, while the AVEC2013 has been validated for affective computing and some mental illnesses, it is possible that feature engineering has some information loss that decreases predictive power.

S3 requires latent embeddings to compute pair-wise scores. Thus, making it suited for deep learning methods. However, applying S3 to traditional ML methods without latent spaces is not straightforward. We choose an S3 variant such that the number of samples selected should equal the number of domains. However, we only investigate this situation in one and two domain settings. Future work will benefit from exploring multi-domain scenarios. Also, notice that selecting more samples with fixed domains will bring the subset closer to the whole dataset, which is undesirable. S3 effectiveness is reduced when transferring from smaller source domain to larger target domains. This is a challenging problem, and future work addressing this area of research is necessary.

S3 is specifically designed with mental health datasets as its focus, so it's best understood within that context rather than as a generic domain adaptation method. Given its effectiveness with smaller datasets and its foundation in speech and signal processing, exploring its use in different areas could be valuable for future research. Such exploration might help determine S3's strengths and weaknesses for general audio data processing.

7.4 Ethical considerations

We believe smartphone speech data for SI detection provides actionable insights for clinicians. However, ethical concerns must be addressed to ensure participant safety and privacy, and our processes are as follows. First, our studies involved at least two clinical psychologists or psychiatrists to address participant concerns. In the MDD study, if suicide intent was expressed (see Fig. 1), a message alerted the team, enabling immediate outreach by a psychiatrist/psychologist. Second, to ensure transparency and accountability when handling sensitive user data, informed consent processes were set. Here, we ensure participants understood how their data was being used through written content and/or custom-made videos. Third, participants are assigned a unique random user ID to protect their identity, and we avoid using demographic data or personally identifiable information (PII) for modeling to minimize unintentional biases and harm. The data is stored on servers with 2FA and only accessible to specific study team members.

The participants recruited in our studies are representative of the psychiatric symptom population. Moreover, we rigorously investigate S3 and other methods under out-of-distribution shifts across four datasets. Nevertheless, users of such systems should be aware of biases in the machine learning models. For instance, our datasets are largely white females. Thus, their effectiveness for minorty groups should be taken into account prior to deployment. We envision our method as a screening tool that works in conjunction with a clinician. The expert will evaluate the detection and suggest appropriate intervention to ensure the individual receives adequate care. Thus, our method is not a substitute for diagnosis or treatment.

7.5 Takeaways and suggestions

Some important findings from our study are as follows:

(1) While most models exhibit poor generalization, models trained on very small datasets benefit from training on larger datasets.
(2) SSDA methods are better suited for the mental health domain as large datasets required for UDA are seldom available.
(3) S3 incrementally improves over SSDA approaches, indicating that a data-centric approach is useful. Nevertheless, generalizability in mental health remains unresolved.

(4) Future studies may benefit from using measures such as effective and relative robustness in addition to accuracy.
(5) HCI and UX researchers should maximize the use of spectral roll-off for design because it is important for SI detection.

8 CONCLUSION

In this paper, we examined the generalizability of speech-based suicidal ideation detection using multiple datasets, including users from different populations. Our analysis indicates that many domain adaptation methods do not generalize well in in-the-wild settings, particularly approaches that require large target datasets. Furthermore, the generalizability of models depends on selecting the "correct" source data for training. Thus, we proposed sinusoidal similarity sub-sampling (S3), which computes pair-wise similarity scores between the source and target domain to select a subset of data for fine-tuning models. Fine-tuning deep models using S3 improves generalizability compared to other deep learning methods across many scenarios. As S3 does not require target labels, it improves generalization considerably on the smallest dataset, suggesting its effectiveness for mental health tasks. In the post-hoc analysis, while two datasets had common important features (spectral roll-off), one dataset had distinct important features, indicating some heterogeneity across populations. Our findings have important practical implications for deploying ubiquitous technology in mental health using machine learning. We hope our work contributes to future research addressing pragmatic challenges in mental health systems, such as distribution shifts, imbalanced datasets, and fine-tuning, ultimately improving mental health screening systems to give individuals the best care possible.

ACKNOWLEDGMENTS

We sincerely thank the reviewers for their valuable comments that significantly enhanced our paper. Our gratitude also extends to the creators of the StudentSADD dataset for their substantial efforts in making such a vital resource available. This research was supported by the National Institute of Mental Health (NIMH) under award numbers R01MH112641 and R01MH123482-01. We acknowledge that the content of this manuscript is solely our responsibility and does not necessarily reflect the views of the National Institutes of Mental Health or the Brain and Behavior Research Foundation. We also clarify that the funding bodies had no involvement in the study's design, data collection, analysis, interpretation, or manuscript preparation.

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Elizabeth L Murnane, Geri Gay, and Tanzeem Choudhury. 2014. Towards circadian computing: " early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 673–684.

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).

[3] Sravanti Addepalli, Samyak Jain, et al. 2022. Efficient and effective augmentation strategy for adversarial training. *Advances in Neural Information Processing Systems* 35 (2022), 1488–1501.

[4] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. 2014. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446* (2014).

[5] David Alvarez-Melis and Nicolo Fusi. 2020. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems* 33 (2020), 21428–21439.

[6] Ronald J Baken. 1987. Clinical measurement of speech and voice. *(No Title)* (1987).

[7] Yuval Barak-Corren, Victor M Castro, Solomon Javitt, Alison G Hoffnagle, Yael Dai, Roy H Perlis, Matthew K Nock, Jordan W Smoller, and Ben Y Reis. 2017. Predicting suicidal behavior from longitudinal electronic health records. *American journal of psychiatry* 174, 2 (2017), 154–162.

[8] Jayme Garcia Arnal Barbedo. 2018. Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification. *Computers and electronics in agriculture* 153 (2018), 46–53.

[9] Anas Belouali, Samir Gupta, Vaibhav Sourirajan, Jiawei Yu, Nathaniel Allen, Adil Alaoui, Mary Ann Dutton, and Matthew J Reinhard. 2021. Acoustic and language analysis of speech for suicidal ideation among US veterans. *BioData mining* 14, 1 (2021), 1–17.

[10] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem Choudhury, Marta Hauser, et al. 2017. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric rehabilitation journal* 40, 3 (2017), 266.

[11] Dror Ben-Zeev, Benjamin Buck, Ayesha Chander, Rachel Brian, Weichen Wang, David Atkins, Carolyn J Brenner, Trevor Cohen, Andrew Campbell, and Jeffrey Munson. 2020. Mobile RDoC: using smartphones to understand the relationship between auditory verbal hallucinations and need for care. *Schizophrenia Bulletin Open* 1, 1 (2020), sgaa060.

[12] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[13] Benjamin Buck, Mary Wingerson, Justin S Tauscher, Matthew Enkema, Weichen Wang, Andrew T Campbell, and Dror Ben-Zeev. 2023. Using smartphones to identify momentary characteristics of persecutory ideation associated with functional disability. *Schizophrenia Bulletin Open* 4, 1 (2023), sgad021.

[14] Thomas Callender and Mihaela van der Schaar. 2023. Automated machine learning as a partner in predictive modelling. *The Lancet Digital Health* 5, 5 (2023), e254–e256.

[15] Scott Carter and Jennifer Mankoff. 2005. When Participants Do the Capturing: The Role of Media in Diary Studies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. Association for Computing Machinery, New York, NY, USA, 899–908. https://doi.org/10.1145/1054972.1055098

[16] Sandeep Nallan Chakravarthula, Md Nasir, Shao-Yen Tseng, Haoqi Li, Tae Jin Park, Brian Baucom, Craig J Bryan, Shrikanth Narayanan, and Panayiotis Georgiou. 2020. Automatic prediction of suicidal risk in military couples using multimodal interaction cues from couples conversations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6539–6543.

[17] Alex Chan, Ahmed Alaa, Zhaozhi Qian, and Mihaela Van Der Schaar. 2020. Unlabelled data improves bayesian uncertainty calibration under covariate shift. In *International Conference on Machine Learning*. PMLR, 1392–1402.

[18] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[19] Kin Wai Cheuk, Kat Agres, and Dorien Herremans. 2020. The impact of audio input representations on neural network based music transcription. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.

[20] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[21] Angélique OJ Cramer, Claudia D Van Borkulo, Erik J Giltay, Han LJ Van Der Maas, Kenneth S Kendler, Marten Scheffer, and Denny Borsboom. 2016. Major depression as a complex dynamic system. *PloS one* 11, 12 (2016), e0167490.

[22] Nicholas Cummins, Julien Epps, Vidhyasaharan Sethu, Michael Breakspear, and Roland Goecke. 2013. Modeling spectral variability for the classification of depressed speech.. In *Interspeech*. 857–861.

[23] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech communication* 71 (2015), 10–49.

[24] Nicholas Cummins, Vidhyasaharan Sethu, Julien Epps, Sebastian Schnieder, and Jarek Krajewski. 2015. Analysis of acoustic space variability in speech affected by depression. *Speech Communication* 75 (2015), 27–49.

[25] Bruce N Cuthbert et al. 2014. The RDoC framework: continuing commentary. *World Psychiatry* 13, 2 (2014), 196.

[26] Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 71–80.

[27] Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2098–2110.

[28] Yizhuo Dong and Xinyu Yang. 2021. A hierarchical depression detection model based on vocal and emotional cues. *Neurocomputing* 441 (2021), 279–290.

[29] Lixin Duan, Dong Xu, and Ivor Tsang. 2012. Learning with augmented features for heterogeneous domain adaptation. *arXiv preprint arXiv:1206.4660* (2012).

[30] Dominic B Dwyer, Peter Falkai, and Nikolaos Koutsouleris. 2018. Machine learning approaches for clinical psychology and psychiatry. *Annual review of clinical psychology* 14 (2018), 91–118.

[31] Omid V Ebrahimi, Julian Burger, Asle Hoffart, and Sverre Urnes Johnson. 2021. Within-and across-day patterns of interplay between depressive symptoms and related psychopathological processes: a dynamic network approach during the COVID-19 pandemic. *BMC medicine* 19, 1 (2021), 1–17.

[32] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[33] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. 2013. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*. 2960–2967.

[34] Judith M Ford. 2016. Studying auditory verbal hallucinations using the RDoC framework. *Psychophysiology* 53, 3 (2016), 298–304.

[35] Daniel Freeman, Emily Bold, Eleanor Chadwick, Kathryn M Taylor, Nicola Collett, Rowan Diamond, Emma Černis, Jessica C Bird, Louise Isham, Ava Forkert, et al. 2019. Suicidal ideation and behaviour in patients with persecutory delusions: Prevalence, symptom associations, and psychological correlates. *Comprehensive Psychiatry* 93 (2019), 41–47.

[36] Daniel Freeman, Bao S Loe, David Kingdon, Helen Startup, Andrew Molodynski, Laina Rosebrock, Poppy Brown, Bryony Sheaves, Felicity Waite, and Jessica C Bird. 2021. The revised Green et al., Paranoid Thoughts Scale (R-GPTS): psychometric properties, severity ranges, and clinical cut-offs. *Psychological Medicine* 51, 2 (2021), 244–253.

[37] Eiko I Fried, Jessica K Flake, and Donald J Robinaugh. 2022. Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology* 1, 6 (2022), 358–368.

[38] Eiko I Fried and Randolph M Nesse. 2015. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR* D study. *Journal of affective disorders* 172 (2015), 96–102.

[39] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[40] John Gideon, Heather T Schatten, Melvin G McInnis, and Emily Mower Provost. 2019. Emotion recognition from natural phone conversations in individuals with and without recent suicidal ideation. In *Interspeech*.

[41] Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems* 17 (2004).

[42] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. 2009. Covariate shift by kernel mean matching. *Dataset shift in machine learning* 3, 4 (2009), 5.

[43] Agnes Gruenerbl, Venet Osmani, Gernot Bahle, Jose C Carrasco, Stefan Oehler, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients. In *Proceedings of the 5th augmented human international conference*. 1–8.

[44] Bonnie Harmer, Sarah Lee, Duong TvH, and Abdolreza Saadabadi. 2020. Suicidal ideation. (2020).

[45] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[46] Jenny Hislop, Sara Arber, Rob Meadows, and Sue Venn. 2005. Narratives of the night: The use of audio diaries in researching sleep. *Sociological Research Online* 10, 4 (2005), 13–25.

[47] Adam Horwitz, Ewa Czyz, Nadia Al-Dajani, Walter Dempsey, Zhuo Zhao, Inbal Nahum-Shani, and Srijan Sen. 2022. Utilizing daily mood diaries and wearable sensor data to predict depression and suicidal ideation among medical interns. *Journal of Affective Disorders* 313 (2022), 1–7.

[48] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398. John Wiley & Sons.

[49] Hsiao-Ying Huang. 2019. Examining reply bias and effectiveness of online community for suicide prevention: A case study of/r/SuicideWatch. In *Social Computing and Social Media. Communication and Social Communities: 11th International Conference, SCSM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31, 2019, Proceedings, Part II 21*. Springer, 108–123.

[50] Parvaneh Janbakhshi, Ina Kodrasi, and Hervé Bourlard. 2020. Subspace-based learning for automatic dysarthric speech detection. *IEEE Signal Processing Letters* 28 (2020), 96–100.

[51] James M Joyce. 2011. Kullback-leibler divergence. In *International encyclopedia of statistical science*. Springer, 720–722.

[52] Ashraf Kagee, Alexander C Tsai, Crick Lund, and Mark Tomlinson. 2013. Screening for common mental disorders in low resource settings: reasons for caution and a way forward. *International health* 5, 1 (2013), 11–14.

[53] Sujay Kakarmath, Andre Esteva, Rima Arnaout, Hugh Harvey, Santosh Kumar, Evan Muse, Feng Dong, Leia Wedlund, and Joseph Kvedar. 2020. Best practices for authors of healthcare-related artificial intelligence manuscripts. *NPJ digital medicine* 3, 1 (2020), 134.

[54] Moslem Khafi, Morteza Fattahi, Hamid Soltanian-Zadeh, and Reza Rostami. 2022. Network-based functional connectivity in MDD with suicide ideation before and after TMS: An fMRI case study. In *2022 30th International Conference on Electrical Engineering (ICEE)*. IEEE, 446–450.

[55] Taekyung Kim and Changick Kim. 2020. Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 591–607.

[56] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2011. Erratum: estimating mutual information [Phys. Rev. E 69, 066138 (2004)]. *Physical Review E* 83, 1 (2011), 019903.

[57] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. , 509–515 pages.

[58] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The PHQ-9: validity of a brief depression severity measure. *Journal of general internal medicine* 16, 9 (2001), 606–613.

[59] Mike Kuniavsky. 2003. *Observing the user experience: a practitioner's guide to user research*. Elsevier.

[60] Eugene Laksana, Tadas Baltrušaitis, Louis-Philippe Morency, and John P Pestian. 2017. Investigating facial behavior indicators of suicidal ideation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 770–777.

[61] Frank Larøi, Iris E Sommer, Jan Dirk Blom, Charles Fernyhough, Dominic H Ffytche, Kenneth Hugdahl, Louise C Johns, Simon McCarthy-Jones, Antonio Preti, Andrea Raballo, et al. 2012. The characteristic features of auditory verbal hallucinations in clinical and nonclinical groups: state-of-the-art overview and future directions. *Schizophrenia bulletin* 38, 4 (2012), 724–733.

[62] Reed Larson and Mihaly Csikszentmihalyi. 2014. The experience sampling method. In *Flow and the foundations of positive psychology*. Springer, 21–34.

[63] Shaokai Li, Peng Song, Keke Zhao, Wenjing Zhang, and Wenming Zheng. 2022. Coupled Discriminant Subspace Alignment for Cross-database Speech Emotion Recognition. *Proc. Interspeech 2022* (2022), 4695–4699.

[64] Weicheng Li, Chengyu Wang, Xiaofeng Lan, Ling Fu, Fan Zhang, Yanxiang Ye, Haiyan Liu, Kai Wu, Yanling Zhou, and Yuping Ning. 2022. Variability and concordance among indices of brain activity in major depressive disorder with suicidal ideation: A temporal dynamics resting-state fMRI analysis. *Journal of Affective Disorders* 319 (2022), 70–78.

[65] Blerina Lika, Kostas Kolomvatsos, and Stathes Hadjiefthymiades. 2014. Facing the cold start problem in recommender systems. *Expert systems with applications* 41, 4 (2014), 2065–2073.

[66] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen. 2010. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering* 58, 3 (2010), 574–586.

[67] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. 2009. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430* (2009).

[68] Alfred Mertins. 1999. *Signal Analysis: Wavelets, Filter Banks, Time-Frequency Transforms and Applications.* John Wiley & Sons.

[69] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2018. A multi-device dataset for urban acoustic scene classification. *arXiv preprint arXiv:1807.09840* (2018).

[70] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13 (2017), 23–47.

[71] Michelle Morales, Stefan Scherer, and Rivka Levitan. 2017. A Cross-modal Review of Indicators for Depression Detection Systems. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology — From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Vancouver, BC, 1–12. https://doi.org/10.18653/v1/W17-3101

[72] Tahsin Mullick, Ana Radovic, Sam Shaaban, Afsaneh Doryab, et al. 2022. Predicting Depression in Adolescents Using Mobile and Wearable Sensors: Multimodal Machine Learning–Based Exploratory Study. *JMIR Formative Research* 6, 6 (2022), e35807.

[73] Humma Nawaz, Ismail Shah, and Sajid Ali. 2023. The amygdala connectivity with depression and suicide ideation with suicide behavior: A meta-analysis of structural MRI, resting-state fMRI and task fMRI. *Progress in neuro-psychopharmacology and biological psychiatry* 124 (2023), 110736.

[74] Matthew D Nemesure, Amanda C Collins, George Price, Tess Z Griffin, Arvind Pillai, Subigya Nepal, Michael V Heinz, Damien Lekkas, Andrew T Campbell, and Nicholas C Jacobson. 2022. Depressive Symptoms as a Heterogeneous and Constantly Evolving Dynamical System: Idiographic Depressive Symptom Networks of Rapid Symptom Changes among Persons with Major Depressive Disorder. https://doi.org/10.31234/osf.io/pf4kc

[75] Alicia L Nobles, Jeffrey J Glenn, Kamran Kowsari, Bethany A Teachman, and Laura E Barnes. 2018. Identification of imminent suicide risk among young adults using text messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.

[76] NSDUH. 2019. SAMSHA Survey 2019. https://www.samhsa.gov/data/sites/default/files/reports/rpt29393/2019NSDUHFFRPDFWHTML/2019NSDUHFFR1PDFW090120.pdf.

[77] Bridianne O'dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions* 2, 2 (2015), 183–188.

[78] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. 2015. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 685–694.

[79] Anastasia Pampouchidou, Olympia Simantiraki, C-M Vazakopoulou, Charikleia Chatzaki, Matthew Pediaditis, Anna Maridaki, Kostas Marias, Panagiotis Simos, Fan Yang, Fabrice Meriaudeau, et al. 2017. Facial geometry and speech analysis for depression detection. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 1433–1436.

[80] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. 2010. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks* 22, 2 (2010), 199–210.

[81] Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak Patel. 2021. Reliable and trustworthy machine learning for health using dataset shift detection. *Advances in Neural Information Processing Systems* 34 (2021), 3043–3056.

[82] Vikram Patel, Digvijay Singh Goel, and Rajnanda Desai. 2009. Scaling up services for mental and neurological disorders in low-resource settings. *International health* 1, 1 (2009), 37–44.

[83] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[84] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. 2019. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1406–1415.

[85] Katharine A Phillips and Rocco D Crino. 2001. Body dysmorphic disorder. *Current opinion in psychiatry* 14, 2 (2001), 113–118.

[86] Kun Qian, Ruolan Huang, Zhihao Bao, Yang Tan, Zhonghao Zhao, Mengkai Sun, Bin Hu, Björn W Schuller, and Yoshiharu Yamamoto. 2023. Detecting somatisation disorder via speech: introducing the Shenzhen Somatisation Speech Corpus. *Intelligent Medicine* (2023).

[87] Branca Telles Ribeiro and Diana de Souza Pinto. 2005. Medical discourse, psychiatric interview. (2005).

[88] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. 2019. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*. 3–12.

[89] Rebecca C Rossom, Karen J Coleman, Brian K Ahmedani, Arne Beck, Eric Johnson, Malia Oliver, and Greg E Simon. 2017. Suicidal ideation reported on the PHQ9 and risk of suicidal behavior across age groups. *Journal of affective disorders* 215 (2017), 77–84.

[90] Daniel M Russell and Ed H Chi. 2014. Looking back: Retrospective study methods for HCI. In *Ways of Knowing in HCI*. Springer, 373–393.

[91] Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. 2019. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8050–8058.

[92] Gary Scavone. 2022. Sinusoidal Dot-Products. Retrieved April 2023 from https://www.music.mcgill.ca/~gary/307/week6/node4.html

[93] Nabeel Seedat, Jonathan Crabbé, and Mihaela van der Schaar. 2022. Data-SUITE: Data-centric identification of in-distribution incongruous examples. In *International Conference on Machine Learning*. PMLR, 19467–19496.

[94] Ankit Parag Shah, Vaibhav Vaibhav, Vasu Sharma, Mahmoud Al Ismail, Jeffrey Girard, and Louis-Philippe Morency. 2019. Multimodal behavioral markers exploring suicidal intent in social media videos. In *2019 International Conference on Multimodal Interaction*. 409–413.

[95] Ankit Singh. 2021. Clda: Contrastive learning for semi-supervised domain adaptation. *Advances in Neural Information Processing Systems* 34 (2021), 5089–5101.

[96] Anjeli Singh and Sareeka Malhotra. 2013. A Researcher's Guide to Running Diary Studies. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction* (Bangalore, India) *(APCHI '13)*. Association for Computing Machinery, New York, NY, USA, 296–300. https://doi.org/10.1145/2525194.2525261

[97] Brian Stasak, Julien Epps, Heather T Schatten, Ivan W Miller, Emily Mower Provost, and Michael F Armey. 2021. Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt. *Speech Communication* 132 (2021), 10–20.

[98] Volkmann Stevens. 1937. Newman, 1937 Stevens SS, Volkmann J., Newman EB. *A scale for the measurement of the psychological magnitude pitch, Journal of the Acoustical Society of America* 8 (1937), 185–190.

[99] Melissa N Stolar, Margaret Lech, Shannon J Stolar, and Nicholas B Allen. 2018. Detection of adolescent depression from speech using optimised spectral roll-off parameters. *Biomedical Journal* 2 (2018), 10.

[100] Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*. Springer, 443–450.

[101] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*. 843–852.

[102] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems* 33 (2020), 18583–18599.

[103] ML Tlachac, Katherine Dixon-Gordon, and Elke Rundensteiner. 2021. Screening for suicidal ideation with text messages. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE, 1–4.

[104] ML Tlachac, Ricardo Flores, Miranda Reisch, Rimsha Kayastha, Nina Taurich, Veronica Melican, Connor Bruneau, Hunter Caouette, Joshua Lovering, Ermal Toto, and Elke A. Rundensteiner. 2022. StudentSADD: Rapid Mobile Depression and Suicidal Ideation Screening of College Students during the Coronavirus Pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 76 (jul 2022), 32 pages. https://doi.org/10.1145/3534604

[105] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7167–7176.

[106] Vihang N Vahia. 2013. Diagnostic and statistical manual of mental disorders 5: A quick glance. *Indian journal of psychiatry* 55, 3 (2013), 220.

[107] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. 2013. Avec 2013: the continuous audio/visual emotion and depression recognition challenge. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*. 3–10.

[108] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[109] Ryan J Van Lieshout and Joel O Goldberg. 2007. Quantifying self-reports of auditory verbal hallucinations in persons with psychosis. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 39, 1 (2007), 73.

[110] Juan Camilo Vásquez-Correa, JR Orozco-Arroyave, T Bocklet, and E Nöth. 2018. Towards an automatic evaluation of the dysarthria level of patients with Parkinson's disease. *Journal of communication disorders* 76 (2018), 21–36.

[111] Colin G Walsh, Jessica D Ribeiro, and Joseph C Franklin. 2017. Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science* 5, 3 (2017), 457–469.

[112] Rui Wang, Min SH Aung, Saeed Abdullah, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A Scherer, et al. 2016. CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 886–897.

[113] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.

[114] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S Kuehn, Jeremy F Huckins, Margaret E Morris, et al. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–34.

[115] Yu-Chu Yu and Hsuan-Tien Lin. 2023. Semi-Supervised Domain Adaptation with Source Label Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24100–24109.

[116] Ansi Zhang, Honglei Wang, Shaobo Li, Yuxin Cui, Zhonghao Liu, Guanci Yang, and Jianjun Hu. 2018. Transfer learning with deep recurrent neural networks for remaining useful life estimation. *Applied Sciences* 8, 12 (2018), 2416.

[117] Yuchen Zhang, Tianle Liu, Mingsheng Long, and Michael Jordan. 2019. Bridging theory and algorithm for domain adaptation. In *International conference on machine learning*. PMLR, 7404–7413.

[118] Zhaowei Zhu, Zihao Dong, and Yang Liu. 2022. Detecting corrupted labels without training a model to predict. In *International conference on machine learning*. PMLR, 27412–27427.

[119] Ayah Zirikly, Philip Resnik, Özlem Uzuner, and Kristy Hollingshead. 2019. CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, Minneapolis, Minnesota, 24–33. https://doi.org/10.18653/v1/W19-3003

A TRADITIONAL MACHINE LEARNING

We perform a parameter search with the following choices: (1) SVM: {kernel:(poly, rbf), C:[0.5, 0.8, 1], gamma:[auto]}, (2) LR: {penalty : [l1, l2], C:np.logspace(-3,3,7), solver:[lbfgs]}, (3) RF: {max_depth: range (2, 10, 1), n_estimators: range(60, 220, 40), min_samples_split: [2, 5, 10], min_samples_leaf: [1, 2, 4]}, and (4) XGB: {max_depth: range (2, 10, 1), n_estimators: range(60, 220, 40), learning_rate: [0.1, 0.01, 0.05]}. The SVM, LR, and RF are implemented using scikit-learn [83], and XGB is implemented using the xgboost [18] package.

A.1 UDA

The UDA apporaches LDM and SA were implemented using the adapt [1] python package. We first perform hyper-parameter tuning as described above. Next, we choose the best model and apply SA or LDM.

B DEEP LEARNING ARCHITECTURES

B.1 Base and S3

We implement Base and VGGish-Z using tensorflow and keras with the architecture shown in Table 9. The models are trained for 500 epochs with a batch size of 32 using the categorical cross entropy loss function with the adam optimizer (lr=1×10^5). Moreover, to prevent overfitting we use earlystopping with a patience=25 and model checkpointing that restores the best model weights. In S3, we fine-tune the models using the best subset for 50 epochs using the same setup as deep learning models with earlystopping and model checkpointing.

B.2 UDA

The UDA approaches MaDD and ADDA were implemented using the adapt [2] python package with Keras and TensorFlow backend. The input to these models is LSTM outputs capturing temporal information from varied length time series. We use the following hyperparameters for the approaches. For MaDD, we used the

[1] https://adapt-python.github.io/adapt/index.html

[2] https://adapt-python.github.io/adapt/index.html

Table 9. VGGish-Z

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, None, 128)]	0
lstm (LSTM)	(None, 128)	131584
dense (Dense)	(None, 128)	16512
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256
dense_2 (Dense)	(None, 2)	130
Total params: 156,482 Trainable params: 156,482 Non-trainable params: 0		

Table 10. Encoder

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 128)	16512
dropout (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 64)	8256
Total params: 24,768 Trainable params: 24,768 Non-trainable params: 0		

Table 11. Predictor Network

Layer (type)	Output Shape	Param #
dense_2 (Dense)	(None, 128)	8320
dense_3 (Dense)	(None, 64)	8256
dense_4 (Dense)	(None, 2)	130
Total params: 16,706 Trainable params: 16,706 Non-trainable params: 0		

encoder (table 10) and predictor (table 11) networks with fully-connected (FC) layers. The model was trained using categorical cross-entropy loss with a batch size 16 and MaDD gamma parameter = 2 for 100 epochs. For optimization, we use stochastic gradient descent with a learning rate (lr) = 0.04 on the encoder and predictor. Furthermore, a learning rate scheduler was applied to reduce lr by one-tenth with momentum and alpha of 0.9 and 0.0002, respectively. Finally, we include early stopping criteria for discriminator loss with patience = 10. ADDA was using the same setup as MaDD. It is worth noting that we implemented the same setup with a 1D ResNet instead of FC layers. However, we did not observe performance improvements.

Table 12. Discriminator

Layer (type)	Output Shape	Param #
dense_5 (Dense)	(None, 128)	8320
dense_6 (Dense)	(None, 64)	8256
dense_7 (Dense)	(None, 2)	130
dense_8 (Dense)	(None, 1)	3
Total params: 16,709 Trainable params: 16,709 Non-trainable params: 0		

B.3 SSDA

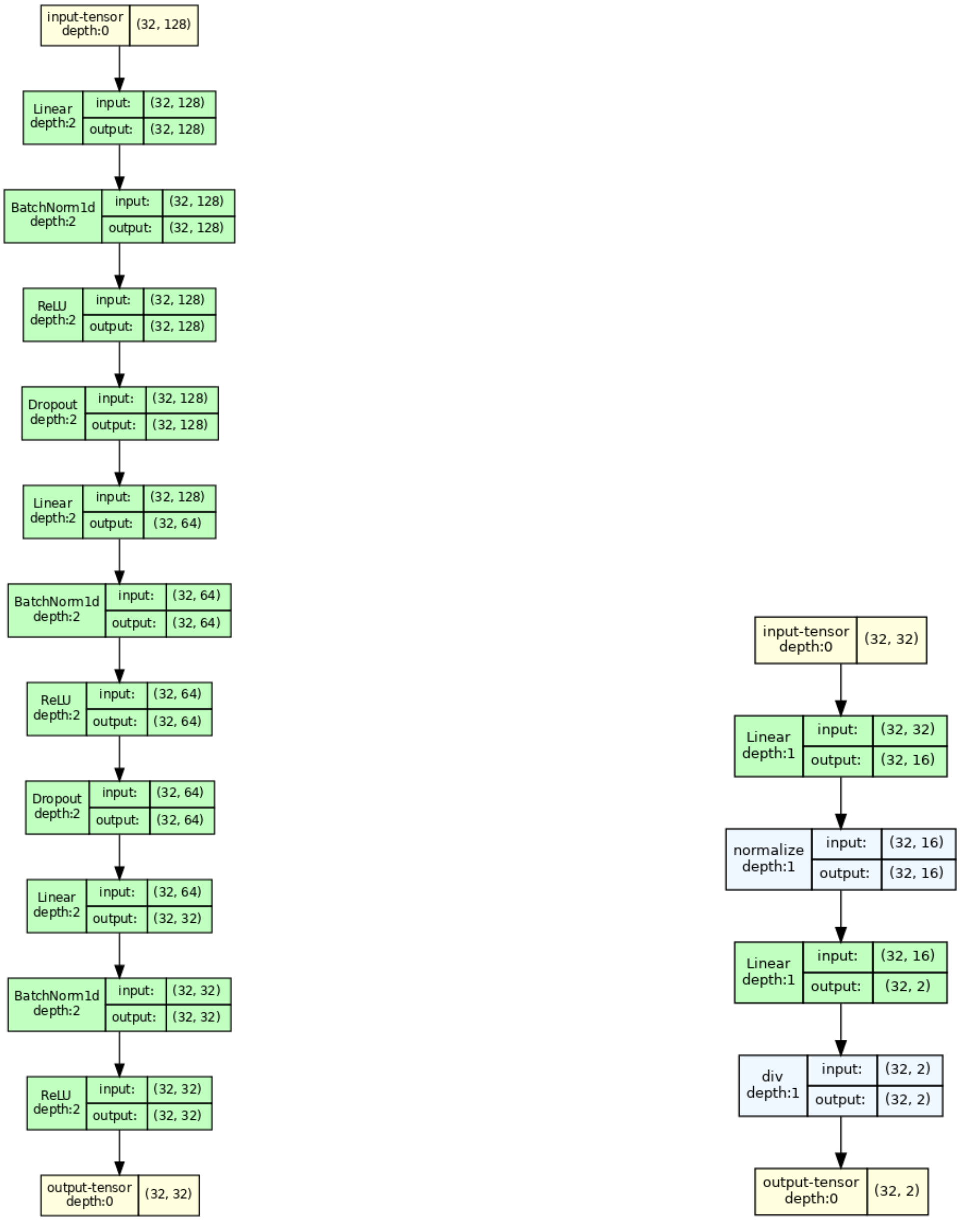
We implemented APE[3], CLDA[4], ENT[5], and MME[6] using PyTorch. Furthermore, it should be noted that CLDA, APE and ENT are built on top of the MME implementation. We use 10% of the target domain as the unlabeled dataset. SSDA consists of two networks, the encoder (Fig. 7a) and the predictor (Fig. 7b). The input to these models is LSTM outputs capturing temporal information from varied length time series. The following hyper-parameters are evaluated on these networks: linear in_features=[128, 64, 32, 16] and dropout p=[0.5, 0.4, 0.3, 0.2]. Overall, the models are trained for 200 epochs with batch size=10 using cross entropy-loss. Furthermore, the G(lr=0.01)and F1(lr=1.0) networks are optimized using stochastic gradient descent with momentum=0.9 and weight decay=0.0005. Now, we discuss specific details for each approach. For ENT and MME, we use temperature=0.05 and eta=0.01. In APE, we extract the normalized output after the first linear layer for computation. Furthermore, we sampled the source data to match the size of target unlabeled data, a requirement for MMD computation. In CLDA, as 2D augmentation cannot be applied to our data, we investigated adding standard gaussian noise and uniform noise. Ultimately, we used uniform noise to generate negative samples for training.

[3] https://github.com/TKKim93/APE

[4] https://github.com/Griffintaur/CLDA_NeurIPS21

[5] https://github.com/VisionLearningGroup/SSDA_MME

[6] https://github.com/VisionLearningGroup/SSDA_MME


input-tensor depth:0 (32, 128)
Linear depth:2 input: (32, 128) output: (32, 128)
BatchNorm1d depth:2 input: (32, 128) output: (32, 128)
ReLU depth:2 input: (32, 128) output: (32, 128)
Dropout depth:2 input: (32, 128) output: (32, 128)
Linear depth:2 input: (32, 128) output: (32, 64)
BatchNorm1d depth:2 input: (32, 64) output: (32, 64)
ReLU depth:2 input: (32, 64) output: (32, 64)
Dropout depth:2 input: (32, 64) output: (32, 64)
Linear depth:2 input: (32, 64) output: (32, 32)
BatchNorm1d depth:2 input: (32, 32) output: (32, 32)
ReLU depth:2 input: (32, 32) output: (32, 32)
output-tensor depth:0 (32, 32)
input-tensor depth:0 (32, 32)
Linear depth:1 input: (32, 32) output: (32, 16)
normalize depth:1 input: (32, 16) output: (32, 16)
Linear depth:1 input: (32, 16) output: (32, 2)
div depth:1 input: (32, 2) output: (32, 2)
output-tensor depth:0 (32, 2)

(a) Encoder (G)

(b) Predictor (F1)

Fig. 7. Encoder and deep latent predictor architectures for SSDA approaches.

C ONE-ONE VALIDATION

C.1 Traditional ML balanced accuracy

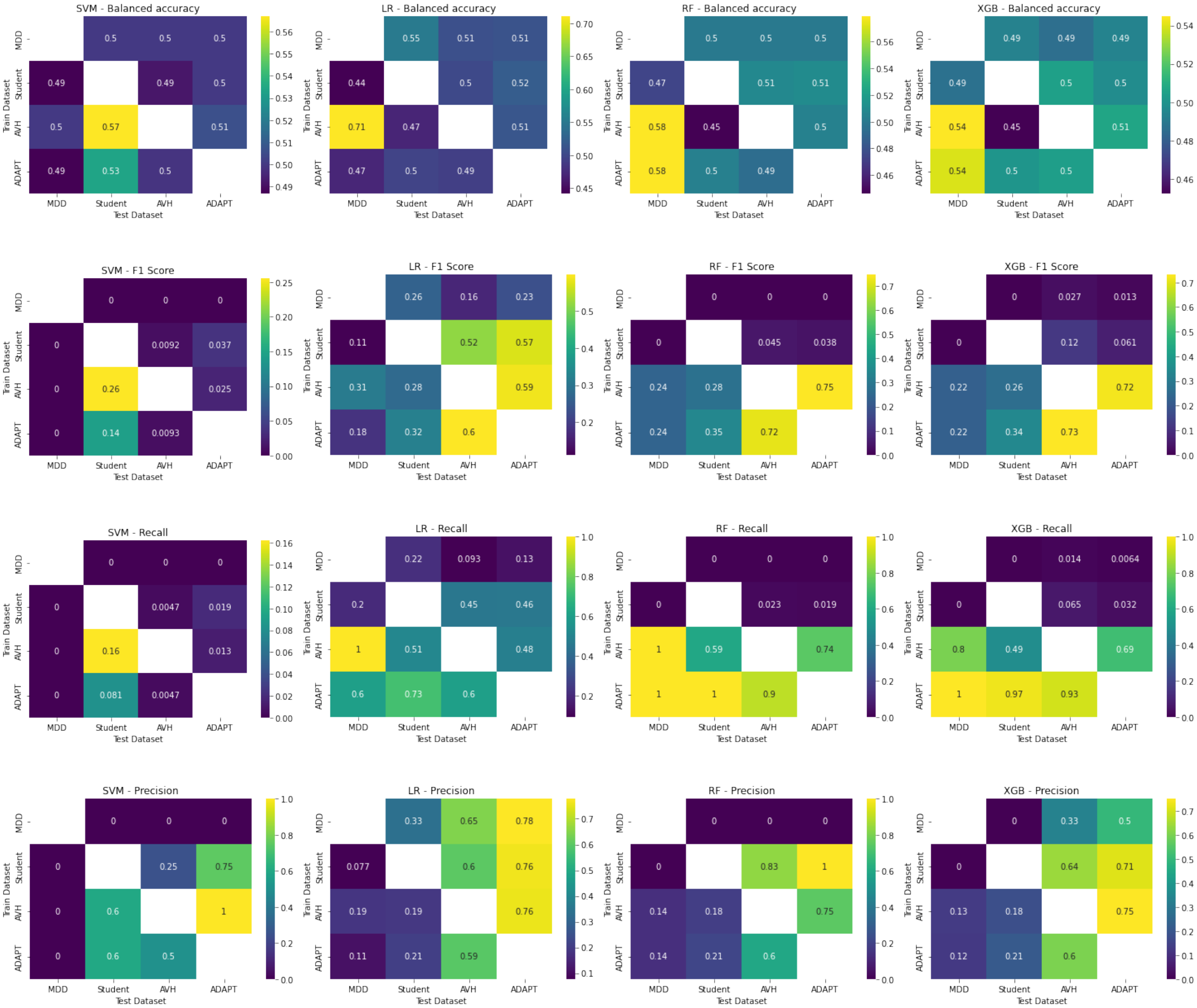
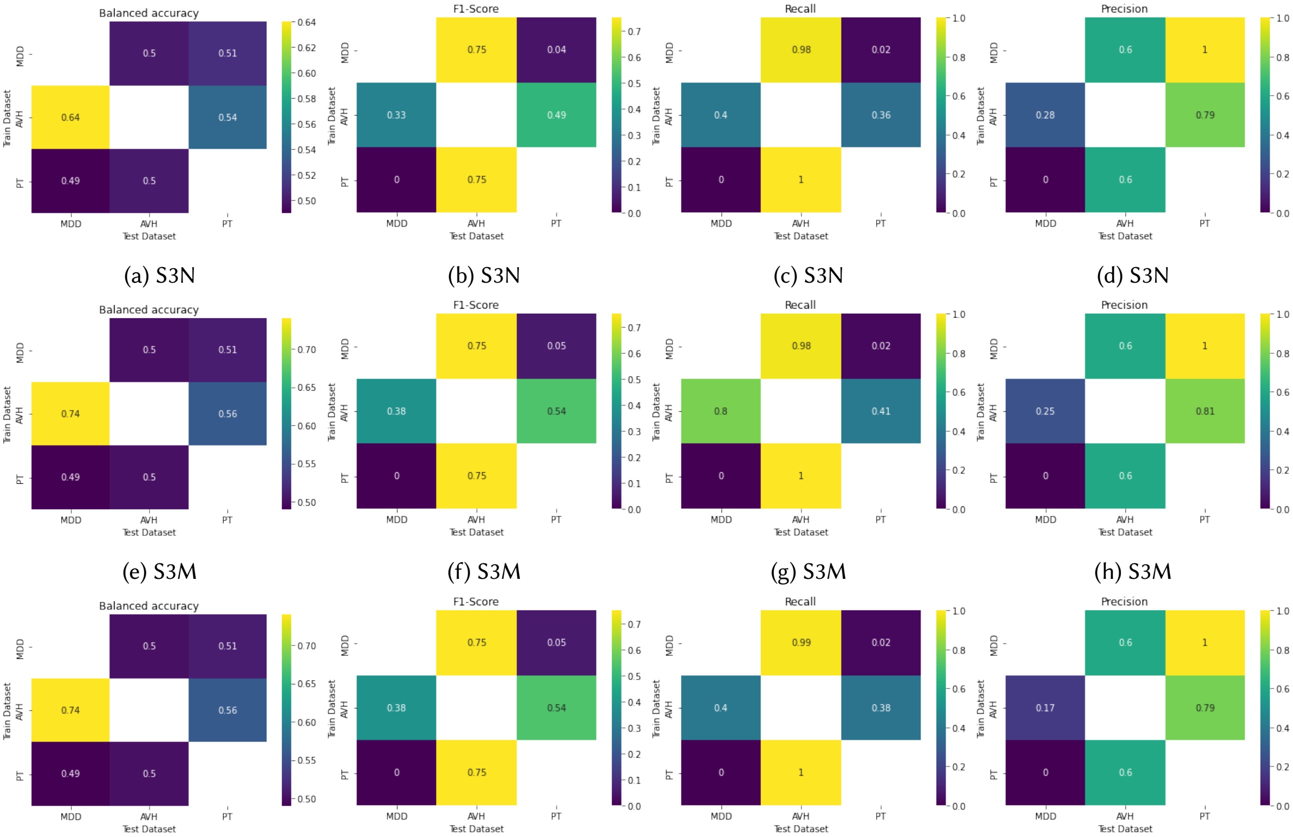


Fig. 8. One-One testing. ADAPT refers to PT.

C.2 S3 variants



(a) S3N (b) S3N (c) S3N (d) S3N

(e) S3M (f) S3M (g) S3M (h) S3M

(i) S3R (j) S3R (k) S3R (l) S3R

Fig. 9. One-One testing.

D DOMAIN ADAPTATION RESULTS

Table 13. Leave-one-dataset-out-validation: domain adaptation (DA) with traditional machine learning methods

DA	**Model**	**Balanced Accuracy**				**F1-Score**			
		MDD	**Student**	**AVH**	**PT**	**MDD**	**Student**	**AVH**	**PT**
LDM	SVM	0.5	NC	0.50	0.47	0.00	NC	0.75	0.62
	LR	0.46	0.54	0.50	0.52	0.16	0.30	0.75	0.71
	RF	0.64	0.54	0.50	0.47	0.27	0.36	0.73	0.65
	XGB	0.50	0.48	0.50	0.50	0.00	0.28	0.00	0.00
SA	SVM	0.50	NC	0.50	0.50	0.21	NC	0.00	0.00
	LR	0.60	0.54	0.46	0.46	0.25	0.3	0.47	0.46
	RF	0.56	0.54	0.50	0.50	0.23	0.36	0.00	0.00
	XGB	0.40	0.48	0.47	0.52	0.09	0.29	0.16	0.06

Table 14. Investigating domain adaptation techniques for acoustic scene classification. Within-Dataset refers to training and testing on the same dataset, whereas all other methods use LODO validation. The three domains A, B, C refer to different recording devices. Metrics are presented as the unweighted mean of the individual class metrics.

Framework	**Method**	**Full Dataset**						**10% of Dataset**					
		Recall			**Precision**			**Recall**			**Precision**		
		A	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**	**A**	**B**	**C**
Within-Dataset	VGGish	0.67	0.49	0.40	0.67	0.49	0.41	0.50	0.27	0.30	0.51	0.13	0.20
LODO	VGGish	0.39	0.24	0.39	0.37	0.37	0.42	0.27	0.31	0.34	0.36	0.31	0.33
	APE [55]	0.18	0.17	0.14	0.04	0.04	0.04	0.17	0.20	0.20	0.04	0.04	0.07
	CLDA [95]	0.35	0.39	0.39	0.34	0.35	0.39	0.21	0.30	0.20	0.06	0.22	0.06
	ENT [41]	0.26	0.23	0.24	0.17	0.24	0.15	0.26	0.30	0.30	0.20	0.16	0.20
	MME [91]	0.37	0.25	0.26	0.26	0.36	0.24	0.29	0.30	0.32	0.31	0.10	0.12
	S3R (proposed)	0.36	0.23	0.36	0.34	0.31	0.39	0.27	0.34	0.33	0.36	0.38	0.39

E ACOUSTIC SCENE CLASSIFICATION

The training procedure for ASC is similar to SI detection. We extracted feature vectors from VGGish to train deep models and SSDA approaches as described in Appendix B and B.3, respectively. The main change is that ASC is a 10-class classification problem, thus, our final linear layer has 10 neurons activated by the softmax function. For within-dataset classification we use 80:20 train-test split within each dataset A, B , and C. They were stratified to maintain the same class distribution. To generate the 10% of Dataset, we randomly chose 10% of the samples in each class. Thus, the full dataset and 10% have the same class distribution. The hyperparameter-tuning procedures used for the SSDA methods are the same as SI detection.