Journal of Affective Disorders 364 (2024) 57–64

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Research paper

Identifying momentary suicidal ideation using machine learning in patients at high-risk for suicide

M.L. Bozzay [a,b,*,1], C.D. Hughes [a,c,1], C. Eickhoff [d], H. Schatten [a,c], M.F. Armey [a,c]

[a] Department of Psychiatry & Human Behavior, Alpert Medical School of Brown University, Box G-BH, Providence, RI 02912, United States
[b] Department of Psychiatry and Behavioral Health, The Ohio State University Wexner Medical Center, 370 W. 9th Avenue, Columbus, OH 43210, United States
[c] Department of Psychosocial Research, Butler Hospital, 345 Blackstone Blvd., Providence, RI 02906, United States
[d] School of Medicine, University of Tübingen, Schaffhausenstr. 77, 72072 Tübingen, Germany

ARTICLE INFO

Keywords:
Ecological momentary assessment
Proximal risk

ABSTRACT

Background: Strategies to detect the presence of suicidal ideation (SI) or characteristics of ideation that indicate marked suicide risk are critically needed to guide interventions and improve care during care transition periods. Some studies indicate that machine learning can be applied to momentary data to improve classification of SI. This study examined whether the classification accuracy of these models varies as a function of type of training data or characteristics of ideation.

Methods: A total of 257 psychiatric inpatients completed a 3-week battery of ecological momentary assessment and measures of suicide risk factors. The accuracy of machine learning models in classifying the presence, duration, or intensity of ideation was compared across models trained on baseline and/or momentary suicide risk data. Relative feature importance metrics were examined to identify the risk factors that were most important for outcome classification.

Results: Models including both baseline and momentary features outperformed models with only one feature type, providing important information in both correctly classifying and differentiating individual characteristics of SI. Models classifying SI presence, duration, and intensity performed similarly.

Limitations: Results of this study may not generalize beyond a high-risk, psychiatric inpatient sample, and additional work is needed to examine temporal ordering of the relationships identified.

Conclusions: Our results support using machine learning approaches for accurate identification of SI characteristics and underscore the importance of understanding the factors that differentiate and drive different characteristics of SI. Expansion of this work can support use of these models to guide intervention strategies.

1. Introduction

Suicide results in 800,000 deaths annually (National Action Alliance for Suicide Prevention, 2014). Patients are at substantially elevated risk of dying by suicide during critical care transitions (Haglund et al., 2019). Intervening early during periods of increases in suicidal ideation (SI) can prevent a cascade to suicidal behavior (SB). However, the onset of SI can occur relatively quickly (Bryan and Rudd, 2016; Kleiman et al., 2017), which makes delivering interventions in a timely manner challenging. Strategies to detect the presence of SI or characteristics SI that indicate marked suicide risk are thus critically needed to guide timely, targeted interventions and improve care during this important transitional period.

Advances in intensive longitudinal sampling, such as Ecological Momentary Assessment (EMA), can facilitate characterization and detection of SI. EMA involves sending brief questionnaires to individuals' mobile phones for completion at different times throughout the day. This method allows for repeated and frequent assessment of experiences as they occur in real-world settings, enabling investigation of short-term changes in both suicide risk processes and in characteristics of SI that can fluctuate over periods of minutes to hours (Kleiman et al., 2017). This approach offers advantages over retrospective self-report approaches, such as providing data with greater ecological validity, and minimizing recall bias (Kendall et al., 1999). Momentary

* Corresponding author at: The Ohio State University Wexner Medical Center, 370 W. 9th Avenue, Columbus, OH 43210, United States.
E-mail address: Melanie.Bozzay@osumc.edu (M.L. Bozzay).
[1] Indicates co-first authorship. Both authors have contributed equally to this manuscript.

https://doi.org/10.1016/j.jad.2024.08.038
Received 15 August 2023; Received in revised form 18 July 2024; Accepted 11 August 2024
Available online 12 August 2024

assessments have minimal social desirability and self-monitoring effects (Hufford et al., 2002), even when assessing suicidality (Coppersmith et al., 2022a). EMA has been increasingly used in the suicide field to describe the onset and trajectory of SI (Kleiman et al., 2017), and to characterize the phenomenological contexts surrounding SI in real-world settings (Armey et al., 2018).

EMA research has provided several key insights about the experience and characteristics of SI that should inform detection strategies. It shows that the phenomenology of SI is highly dynamic, with the presence and intensity of SI often fluctuating over periods of hours to days (Bryan and Rudd, 2016; Kleiman et al., 2017). While SI has usually been studied as a homogenous construct, understanding the characteristics of SI may help to distinguish which patients with SI are most likely to make a future suicide attempt (Bryan et al., 2019). Although work in this area is in its early stages, suicide risk profiles are distinguishable by characteristics of SI, such as how intensely and frequently SI occurs (Bryan et al., 2019). These findings underscore the clinical relevance of characterizing the phenomenology surrounding not only the presence of SI, but also characteristics of SI such as its intensity and duration.

Several factors have been linked with risk of SI in general. Characteristics typically measured during initial, or *baseline*, patient assessments such as psychiatric diagnoses, suicide risk history, and facets of emotional (i.e., hopelessness), cognitive (i.e., rumination), and behavioral (e.g., impulsivity) functioning have been linked with current and subsequent SI (Allen et al., 2019; Franklin et al., 2017). However, associations between these "baseline" risk factors and SI tend to be small (Franklin et al., 2017), reflecting the challenge of using historical factors captured at a single point in time to predict an outcome that may be highly variable and influenced by ecological or contextual factors. Recently, there has been an emphasis on identifying *momentary* factors occurring close in time to SI that may signify the presence of suicide risk states (Galynker et al., 2017). Studies show that momentary risk factors measured via EMA at the state level are short-term correlates of the onset of momentary ideation (Armey et al., 2018; Kleiman et al., 2017). However, whether these factors have similar or differential utility in accurately classifying characteristics of SI (i.e., intensity, duration) in vivo is unclear.

Nevertheless, the complex nature of suicide risk has historically made the development of models to accurately classify characteristics of suicide risk challenging. Suicide theory suggests that relationships between risk factors and SI are nonlinear and dynamic (Bryan and Rudd, 2016), and there are likely complex interrelationships between risk factors that are important for more accurately identifying suicide risk. These complexities are challenging to model via traditional regression-based approaches which require pre-specification of these relationships and parameters. Advanced modeling approaches such as machine learning are data-driven and uniquely designed to maximize model accuracy by handling large volumes of variables and intensively modeling complex associations between classifiers and outcome(s). Some meta-analyses show that these data-driven ML approaches are highly promising and can predict SI and SB with up to an 18-fold higher odds ratio than the theoretical, regression-based models that historically have been used in the field (Schafer et al., 2021). Importantly, however, several studies cited in these meta-analyses have since been criticized, or even retracted due to failing to properly validate their ML models, resulting in overfitting and potentially overemphasizing the benefits of ML in suicide risk classification efforts (Jacobucci et al., 2021; Just et al., 2023). As a result, existing research elucidating the potential relative benefits of ML-based approaches over other simpler approaches (i.e., generalized linear models and other forms of linear models) is limited.

Moreover, most studies using ML have used information about baseline patient characteristics (e.g., psychiatric history) to identify which patients are at increased risk of SI months to years later (Schafer et al., 2021). Only a few studies have successfully applied ML to classify suicide risk from momentary data (i.e. risk factors, features of ideation), but those that did showed that next-day SI in youth (Czyz et al., 2021a) and near-term SB in adults (Wang et al., 2021) can be predicted from these data. While these studies suggest that ML approaches using momentary data can improve prediction or classification of SI/SB, they have important limitations. First, despite research indicating that there are likely both longstanding/stable and time-varying/dynamic aspects of suicide risk (Bryan and Rudd, 2016), it is unknown if and how the classification accuracy of short-term SI risk varies across models trained on different types of data (i.e., baseline, momentary, or both). Second, which risk factors are *most* important in accurately classifying patients at increased short-term risk of SI is unclear. Third, whether model accuracy and/or variable importance differ when classifying different characteristics of SI (i.e., presence, duration, intensity) is unknown. Fourth, whether the use of ML models to accomplish these aims is superior to simpler models is unknown. This study will examine these important research questions to derive information necessary to develop more accurate risk classification models and targeted and timely intervention strategies.

2. Methods

2.1. Participants

Participants were 460 patients at an inpatient psychiatric hospital in the northeastern United States. We recruited inpatients hospitalized for SI (72 %), SA (15 %), and no history of SA and no SI in the month prior to hospitalization (13 %). Inclusion criteria were aged 18–70, English fluency, and comfort with smartphones. Current psychotic/manic symptoms severe enough to interfere with participation were exclusionary. Analyses included only participants who completed EMA (n = 257). Participants varied in age (M = 40.53, SD = 13.33), with 54 % women. Approximately 88 % of participants were white, 6 % Black/African American, 2 % Asian, 1 % as American Indian/Native American. Most were non-Hispanic (91 %). Most were single/never married (45 %) or divorced/separated (25 %).

2.2. Procedures

Staff screened patient charts for eligibility. Patients provided informed consent and completed an assessment battery to ascertain eligibility and measure SI risk factors. Interviews were administered by bachelor's level staff supervised by a licensed clinical psychologist. Following discharge, participants received EMA prompts to complete brief (<5 min) assessments scheduled four times a day, at least 1 h apart, at random intervals over three-weeks. Participants also completed identical, self-initiated assessments during times when they engaged in suicidal or non-suicidal self-injurious behavior or experienced an exacerbation in SI. Study procedures were approved by the Butler Hospital IRB.

2.3. Measures

2.3.1. Baseline risk factors

2.3.1.1. Depressive symptoms. The Quick Inventory of Depressive Symptomatology assesses the severity of depressive symptoms over the past week (Rush et al., 2003).

2.3.1.2. Borderline personality disorder symptoms. The McLean Screening Instrument for Borderline Personality Disorder (Zanarini et al., 2003) screens for symptoms of borderline personality disorder (Cronbach's alpha = 0.76).

2.3.1.3. Negative attitudes. The Dysfunctional Attitudes Scale (DAS) (Weissman and Beck, 1978) measures pervasive negative attitudes towards the self, the world, and the future (Cronbach's alpha = 0.94).

2.3.1.4. Childhood trauma. The Childhood Trauma Questionnaire (Bernstein et al., 1994) assesses the severity of different types of childhood trauma (Cronbach's alpha ranged from 0.75 to 0.95 across subscales).

2.3.1.5. Impulsivity. The Barrett Impulsiveness Scale (Patton et al., 1995) assesses different facets of impulsive tendencies (Cronbach's alpha = 0.65–0.73).

2.3.1.6. Emotional dysregulation. Trait-level perceived ability to regulate emotions was assessed using the 36-item Difficulties in Emotion Regulation Scale (DERS) (Gratz and Roemer, 2004) (Cronbach's alpha ranged from 0.77 to 0.90 across subscales).

2.3.1.7. Acquired capability. The Acquired Capability for Suicide Scale (Van Orden et al., 2008) (ACSS) assesses fearlessness of death and perceived tolerance for physical pain (Cronbach's alpha = 0.33).

2.3.1.8. Perceived burdensomeness and thwarted belongingness. The Interpersonal Needs Questionnaire (Van Orden et al., 2012) measures perceptions of burdensomeness and low belongingness (Cronbach's alphas = 0.87–0.91).

2.3.1.9. Depressive rumination. Tendencies towards brooding and pondering depressive rumination were assessed using subscales from the Response Styles Questionnaire (RSQ) (Nolen-Hoeksema and Morrow, 1991) (Cronbach's alphas = 0.65–0.80).

2.3.1.10. Hopelessness. We used the Beck Hopelessness Scale (Beck, 1988) to assess negative expectations for the future (Cronbach's alpha = 0.91).

2.3.1.11. Suicide attempt history. We assessed lifetime frequency of suicide attempts using the Columbia Suicide Severity Rating Scale (C-SSRS) interview (Posner et al., 2008).

2.3.2. Momentary risk factors

Momentary risk factors were assessed via EMA prompts delivered via Ilumivu's HIPAA certified mEMA phone application, which provides a cross-platform (iOS and Android) application for delivery of multiple simultaneous scheduled EMA protocols. Participants completed an average of 33 (SD = 31.18) EMA surveys, resulting in 8412 completed surveys. SI was endorsed in 1043 (13.10 %) EMAs.

2.3.2.1. Response context. Participants reported their location, whether they were alone, and whether they had used substances since the last assessment.

2.3.2.2. Negative life events. Participants reported whether they had experienced a negative event since the last assessment.

2.3.2.3. Positive and negative affect. Items measuring positive (e.g., "happy") and negative affect (e.g., "sad") were derived from the PANAS-X (Watson and Clark, 1994).

2.3.2.4. Ruminative thinking and emotional reactivity. Items assessed current difficulties in emotional regulation from the DERS (Gratz and Roemer, 2004), and ruminative tendencies from the RSQ (Nolen-Hoeksema and Morrow, 1991).

2.3.2.5. Distress tolerance. Participants answered items pertaining to their ability to manage distress from the Distress Tolerance Scale (Simons and Gaher, 2005).

2.3.2.6. Non-suicidal self-injury. Participants reported non-suicidal self-injury since the last assessment.

2.3.3. Momentary outcomes

2.3.3.1. Suicidal ideation characteristics. Items based on the Modified Scale for Suicide Ideation (Miller et al., 1986) assessed the presence, duration [Shorter: SI denied - several minutes vs. Longer: an hour or more –continuously], or intensity [Lower: SI denied-weak vs. Higher: strong-very strong] of SI since last assessment.

2.3.4. Data analytic strategy

2.3.4.1. Random forest algorithm. We used random forest (RF) classifiers (Breiman, 2001) to model the data.[2] RF models are classification algorithms made up of ensembles of decision trees. Decision trees model the relationships among predictor variables and outcomes as a series of nodes and splits/branches, where each node uses one variable to make a separating decision, or split the data to optimally partition classes, which when compounded over several nodes/branches provides probabilities for each classification of new data. In RFs, each tree is trained using a different bootstrap sample of the training data (i.e., unique datasets generated by randomly resampling the training dataset) containing a randomly selected subset of all available predictor variables. To predict the classification of new data, each tree 'votes' for one class and the RF selects the class with a majority of votes. Specific rules for tree growing, tree combination, and self-testing make RF models robust to overfitting, outliers, and noisy data, and well-suited to non-linear relationships and high-dimensional data (Caruana and Niculescu-Mizil, 2006; Menze et al., 2009). RFs are built, in part, by evaluating the importance of variables based on their Gini importance, an importance score for each predictor variable based on the frequency the variable was used to make a decision, weighted by the number of samples it classifies, and averaged across all ensemble trees, which can also be used to rank the importance of each predictor variable (Colic et al., 2022).

2.3.4.2. Missing data. A special reserved value of −1 was assigned to missing observations. As there were no organically observed negative values, this negative value serves as a special flag that allows the model to explicitly discern missing data and reason over non-random patterns of missingness. In this particular application (as opposed to say vital sign features in the electronic health record), we do not expect data to be missing at random, but rather following an underlying trend. That is, the fact that an EMA or other item was not answered might hold information about the participant's situational mental state. It was for this reason that we introduced a separate reserved feature value to indicate missingness rather than hiding it from the ML algorithm via traditional (e.g., mean) imputation.

2.3.4.3. Description of model training conditions. We trained and compared a separate model for each of three different momentary SI characteristics outcomes (SI presence, intensity, and duration rated during EMAs) to assess and compare model performance when different types of data were included in the models. Specifically, we trained three models using the demographic (i.e., age, sex, and number of lifetime SAs) with 1) baseline risk factors (e.g., hopelessness, emotional dysregulation, and suicide attempt history), 2) momentary risk factors (e.g., positive and negative affect, and rumination), or 3) baseline risk factors, *and* momentary variables. These three models were trained to classify

[2] The unit of classification in this study (i.e., a training or test data point) were individual EMA surveys, of which participants completed an average of 33 (SD = 31.18), providing 8412 data points. These data points provide a sufficient sample size for present analyses.

each of our three SI characteristic outcome variables: SI presence/absence (Models A1, A2, and A3), SI duration (Models B1–B3), and SI intensity (Models C1–C3), resulting in nine RF models.

2.3.4.4. Model training. We used a stratified 80/20 split for training/test data, where 80 % of the data was available to the model for training, and 20 % remained unknown to the model and was used to evaluate model performance and generalization. A 10-fold cross-validation (no repeats, stratified folds) of the training set was used for hyperparameter tuning. A grid search method was used to tune each of the hyperparameters for the RF models: (1) number of trees in the forest; (2) number of splits within each tree; and (3) minimum number of data points per split.

2.3.4.5. Class imbalance. Like many clinical problems, our data shows a stark class imbalance skewed towards the negative class that can sometimes impede the success of machine learning models. To counter these detrimental effects, we applied random oversampling via SMOTE (Chawla et al., 2002) to obtain a 1:1 class balance in the training data. The class distribution of the test set was not changed and remains at the original clinical incidence rate.

2.3.4.6. Model evaluation. After model training, models were asked to classify new observations (the previously unseen 20 % of the data) and the model performance was determined by calculating the accuracy, precision, recall/sensitivity, specificity, negative predictive value (NPV) and the Area Under the Curve (AUC) of the Receiver Operator Coefficient (ROC). Results are also visually depicted via confusion matrices (Fig. 2).

2.3.4.6.1. Variable importance. The Gini importance was used to rank order the importance of variables/factors in each model and compare them across models.

2.3.4.7. Additional comparisons. We also conducted additional analyses in which we supplemented the performance part of the train set to examine the acceptability of the degree of overfitting or underfitting of the model. We also ran additional models (i.e., GLM, LDA, and ElasticNet) to compare these random forest models to that of simpler, linear ML models. Details of these experiments are provided in the Supplement, and the performance of these different models is presented in Table 1. All experiments were conducted in Python, using the numpy (linear algebra), scikit-learn (machine learning), imblearn (resampling), and matplotlib (plotting) libraries. The tuning parameter free GLM and LDA were added. For the ElasticNet, we cross-validated alpha (overall regularization strength, 0.0–5, increments of 0.1) and the L1 ratio (mixing parameter controlling the relative importance of L1 vs. L2 priors, 0.0–1.0, increments of 0.1), the number of iterations (100–1000, increments of 100), and whether or not to estimate the intercept (yes/no). The optimal model whose performance is reported here ended up using alpha = 1.3, L1-ratio = 1.0, iterations = 800, estimate intercept = True.

3. Results

3.1. Model performance

The accuracy, precision, recall/sensitivity, specificity, NPV, and ROC-AUC for models are in Table 1. Overall, model performance improved as more data/variables were made available for training. Models trained with baseline and EMA variables generally outperformed models trained using either baseline or EMA variables, with some nuance. For all three SI characteristics, the recall/sensitivity improved for models trained on both EMA and baseline data, however, models trained on baseline data alone produced higher recall/sensitivity than that of models trained on both baseline and EMA data, indicating limited

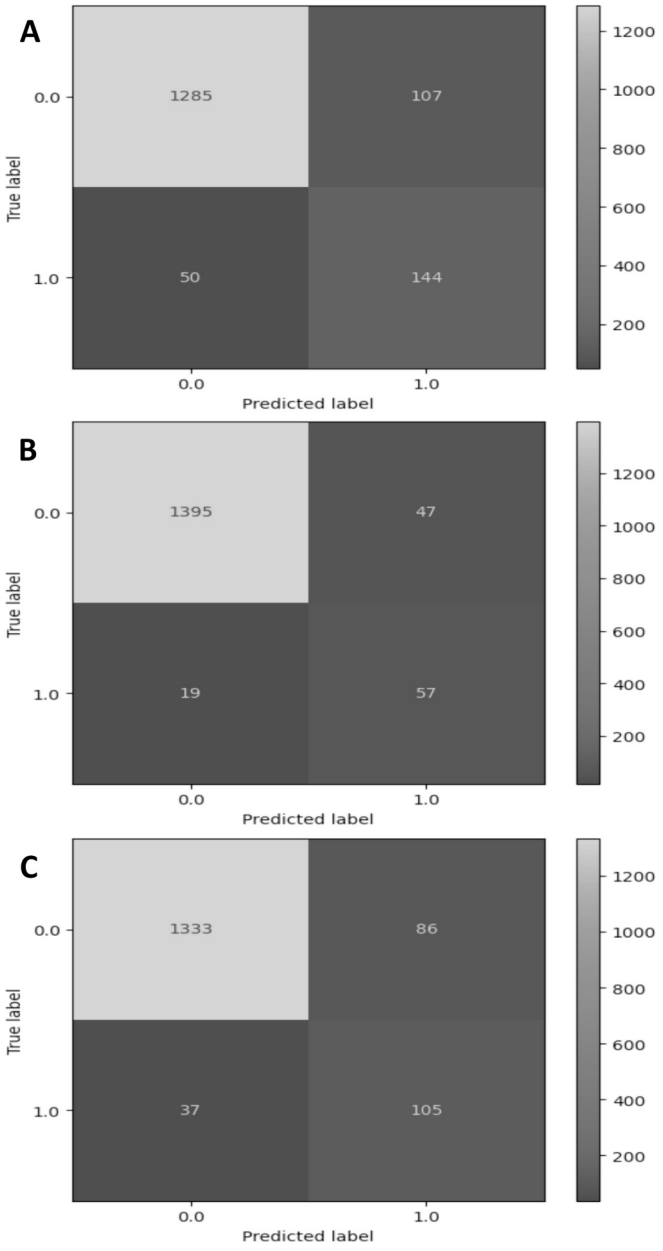

Fig. 2. Confusion matrices for models classifying A) SI Presence (3A), B) SI Duration (3B), and C) SI Intensity (3C).

improvement in false negative rates. Models 3A–C (Table 1) were the best performing across outcomes (while recall/sensitivity was not optimal in best performing models, all other metrics were) with comparable performance to one another on all metrics (accuracy, precision, recall/sensitivity, specificity, NPV, and ROC-AUC). The best tuning parameter values were: (1) number of trees in the forest = 100; (2) maximum number of splits per tree = 8; and (3) minimum number of data points per split = 2. These hyperparameters were used consistently across all predictive endpoints and datasets.

3.2. Important features

The 20 most important features (by Gini importance) for Models 3A–C are presented in Fig. 1 and Table 2, and discussed below. All models shared the same top five important features (momentary hopelessness, sadness, experiencing emotions as overwhelming, having difficulty making sense of feelings, and thinking "why do I always react this

Table 1
Models' performance by classification target.

		a. Presence						b. Duration						c. Intensity					
		AUC	Rec	Spec	Prec	NPV	Acc	AUC	Rec	Spec	Prec	NPV	Acc	AUC	Rec	Spec	Prec	NPV	Acc
1) BL	GLM	0.68	0.73	0.63	0.22	0.93	0.64	0.70	0.71	0.70	0.11	0.98	0.70	0.72	0.76	0.69	0.19	0.96	0.69
	LDA	0.68	0.71	0.65	0.22	0.93	0.66	0.72	0.71	0.72	0.12	0.97	0.72	0.74	0.77	0.70	0.20	0.96	0.71
	EN	0.68	0.71	0.64	0.22	0.93	0.65	0.71	0.71	0.71	0.12	0.98	0.71	0.73	0.75	0.70	0.20	0.96	0.70
	RF	0.80	0.79	0.80	0.36	0.95	0.80	0.84	0.80	0.88	0.25	0.99	0.87	0.82	0.80	0.83	0.32	0.97	0.83
2) EMA	GLM	0.75	0.71	0.80	0.33	0.95	0.78	0.79	0.74	0.84	0.19	0.98	0.83	0.77	0.74	0.81	0.28	0.97	0.80
	LDA	0.75	0.73	0.77	0.31	0.95	0.77	0.79	0.78	0.81	0.18	0.99	0.81	0.78	0.76	0.80	0.27	0.97	0.80
	EN	0.76	0.73	0.79	0.32	0.95	0.78	0.79	0.76	0.82	0.18	0.98	0.82	0.77	0.73	0.81	0.28	0.96	0.80
	RF	0.80	0.68	0.92	0.55	0.95	0.89	0.81	0.66	0.96	0.49	0.98	0.95	0.80	0.65	0.94	0.52	0.96	0.91
3) BL + EMA	GLM	0.76	0.72	0.81	0.35	0.95	0.80	0.82	0.79	0.85	0.21	0.99	0.84	0.80	0.77	0.83	0.31	0.97	0.82
	LDA	0.77	0.74	0.81	0.35	0.95	0.80	0.82	0.80	0.84	0.21	0.99	0.84	0.80	0.79	0.82	0.30	0.97	0.82
	EN	0.77	0.74	0.80	0.34	0.95	0.80	0.81	0.78	0.84	0.21	0.99	0.84	0.80	0.77	0.83	0.32	0.97	0.83
	RF	0.83	0.74	0.92	0.57	0.96	0.90	0.86	0.75	0.97	0.55	0.99	0.96	0.84	0.74	0.94	0.55	0.97	0.92

EMA = Ecological Momentary Assessment; BL = Baseline; GLM = Generalized Linear Model; LDA = Linear Discriminant Analysis; EN = Elastic Net; RF = Random Forest; Acc = Accuracy; AUC = Area Under the Curve; Prec = Precision; Rec = Recall/sensitivity; Spec = Specificity; NPV = Negative Predictive Value.

way?"). Twelve features were within the top 20 of all three models, 11 of which were EMA-measured, and one was baseline-assessed (depression severity).

3.3. Additional model comparisons

When compared with RF models, other modeling approaches performed comparably or worse on all metrics except recall/sensitivity (see Table 1). The superior recall/sensitivity of LDA models was offset by their worse performance on all other metrics, with the relatively poor precision particularly undercutting LDA utility due to elevated rates of false positives. Moreover, the alternative modeling approaches slightly outperformed RF models in recall/sensitivity, indicating slightly lower false negative rates.

4. Discussion

In this study, we applied machine learning methods to baseline and momentary risk factor data to classify the presence, duration, and intensity of momentary SI. To our knowledge, this is the first study to evaluate the utility of both baseline *and* momentary risk factors in machine learning models classifying different characteristics of momentary SI.

Regardless of SI outcome, models trained with *only* baseline or *only* momentary data were outperformed by models trained using both sources. This is consistent with prior EMA research which better predicted SI compared to models trained with fewer variables (Czyz et al., 2021a; Czyz et al., 2021b; Wang et al., 2021). Interestingly, we found that models trained with *only* momentary data tended to do better than models trained with only baseline data, which makes sense given that our outcomes were momentary in nature.

Assessing and comparing the relative importance of variables/factors across SI outcomes (Models 3A-C) yielded a nuanced pattern of results. For all models, momentary hopelessness was the most important classifier of SI. All models shared their top five variables—albeit in different orders—which were constructs reflecting momentary negative affect (i.e., hopelessness and sadness) and momentary emotion reactivity (i.e., experiencing emotions as overwhelming, having difficulty making sense of emotions, and thinking "why do I always react this way?"). This is in line with existing models of suicidality (Bryan and Rudd, 2016; Selby and Joiner, 2013) and research about the risk processes that indicate acute suicidal crises (Galynker et al., 2017). Surprisingly, constructs frequently studied in relation to suicide risk, including borderline personality disorder symptoms and interpersonal theory-related factors were not among the 20 most important factors in any of the models examined. These findings are consistent with research suggesting that these constructs are useful indicators of longstanding risk for suicide but are not markers of shifts into elevated suicidal risk states (Galynker et al., 2017). That most contextual variables (e.g., location) assessed via EMA were not among the top classification features suggests the internal context (i.e., affect and cognitions) is more relevant to identifying characteristics of SI.

Some differences emerged when top features of models were compared. Models all had the same top five classifying variables and shared eight of the top ten variables, which indexed affective, emotional dysregulation, and momentary ruminative thinking features previously implicated in momentary SI (Armey et al., 2018). In contrast, childhood maltreatment (emotional and sexual abuse) seemed relevant to SI intensity and duration but not presence. Additionally, variables pertaining to longstanding lack of emotional clarity, trait attentional impulsivity, and momentary/EMA-rated confidence differentiated the model indexing SI duration. These findings highlight the relevance of trait-like indicators, suggesting the intensity and duration of SI are more related to dispositional, memory, or self-evaluative factors compared to SI presence. These results are consistent with research underscoring variation in the patterns and factors associated with different characteristics of SI (Coppersmith et al., 2022b).

This diverging pattern of results across models also underscores the importance of studying both baseline and momentary risk factors in relation to state-level manifestations of a range of characteristics of SI. Across all models, 4–6 of the 20 most important features were baseline-assessed constructs. However, the most important five features for each model were EMA constructs. Taken together with the superior performance from the combined models (3A–C), these findings seem to suggest that while momentary constructs are more robust classifiers of SI, baseline/stable constructs are still relevant especially for distinguishing different characteristics of SI. Therefore, effective characterization of aspects of SI as well as the best classification across all SI outcomes appears to result from their combined use.

There were also important differences in performance of the various modeling approaches (i.e., RF models compared to other modeling approaches). In general, we found that RF models outperformed other modeling approaches on most metrics (accuracy, precision, specificity, NPV, & AUC, indicating better false-positive and true-positive rates. Notably, however, we also found that RF models' recall/sensitivity was comparable or slightly worse than other modeling approaches, suggesting that our RF models were slightly more prone to misclassify individuals as not having suicidal ideation. Given the potential deleterious costs of misclassification for suicidal individuals—both false positive and false negatives—efforts to implement these kinds of models in clinical settings should carefully consider the relative cost/benefit of bias towards false-positive or false-negatives given the context of the application, and future research in this area should pay particular attention to model sensitivity and consider additional strategies to

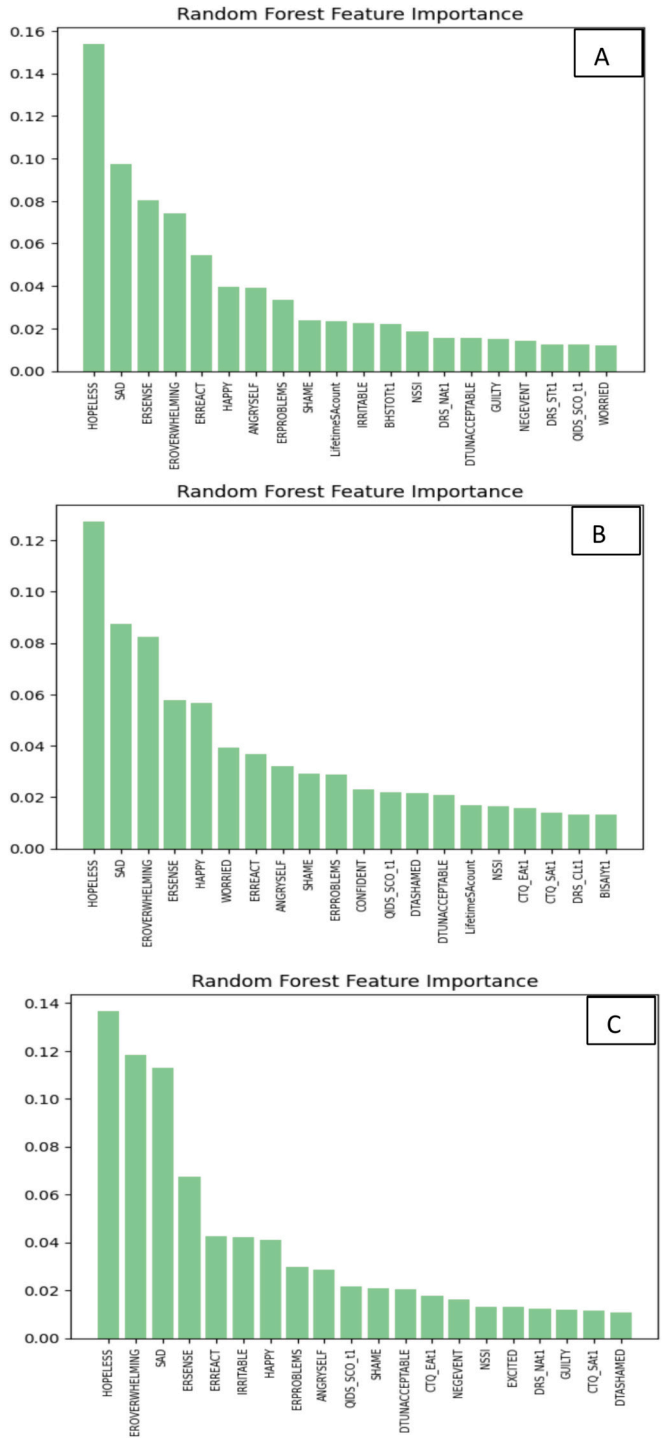

Fig. 1. Random forest feature importance for models classifying A) SI Presence (3A), B) SI Duration (3B), and C) SI Intensity (3C).

Table 2
Importance of features included in models.

Feature	Presence	Duration	Intensity
EMA-rated Hopelessness	1	1	1
EMA-rated Sadness	2	2	3
Having Difficulties Making Sense of Feelings	3	4	4
Experiencing emotions as overwhelming	4	3	2
Thinking "Why do I always react this way?"	5	4	5
EMA-rated Happiness	6	5	7
EMA-rated Anger at Self	7	8	9
Thinking "Why do I have problems others don't?"	8	10	8
EMA-rated Shame	9	9	11
Lifetime SA Count	10	15	>20
EMA-rated Irritability	11	>20	6
Beck-rated Hopelessness	12	>20	>20
Endorsement of NSSI (yes/no)	13	16	15
DERS Non-acceptance of emotions	14	>20	17
Experiencing Distress as Unacceptable	15	14	12
EMA-rated Guilt	16	>20	18
Negative Event (yes/no)	17	>20	14
DERS Limited Access to Emotion Regulation Strategies	18	>20	>20
QIDS-rated Depression Level	19	12	10
EMA-rated Worry	20	6	>20
Experiencing Shame Regarding Own Distress	>20	13	20
Childhood Emotional Abuse	>20	17	13
Childhood Sexual Abuse	>20	18	19
EMA-rated Confidence	>20	11	>20
DERS Lack of Emotional Clarity	>20	19	>20
BIS Attentional Impulsivity	>20	20	>20
EMA-rated Excitement	>20	>20	10
Age	>20	>20	>20
Sex/Gender	>20	>20	>20
Time Since Discharge	>20	>20	>20
RSQ Brooding Subscale	>20	>20	>20
RSQ Pondering Subscale	>20	>20	>20
DERS Lack of Emotional Awareness	>20	>20	>20
DERS Difficulties with Goal-Directed Behavior	>20	>20	>20
DERS Impulse Control Difficulties	>20	>20	>20
BIS Impulsive non-planning	>20	>20	>20
BIS Motor Impulsivity	>20	>20	>20
Childhood Physical Abuse	>20	>20	>20
Childhood Physical Neglect	>20	>20	>20
Childhood Emotional Neglect	>20	>20	>20
Childhood Positive Family Score	>20	>20	>20
BPD Features	>20	>20	>20
Dysfunctional Attitudes	>20	>20	>20
Acquired Capability for Suicide	>20	>20	>20
Thwarted Belongingness	>20	>20	>20
Perceived Burdensomeness	>20	>20	>20
EMA response type (random vs user initiated)	>20	>20	>20
Current Location	>20	>20	>20
Isolation (alone vs with others)	>20	>20	>20
Substance Use (yes/no)	>20	>20	>20
Thinking about recent situation and wishing it had gone better	>20	>20	>20
Desire to Avoid Feeling Distressed	>20	>20	>20
Experiencing Distress as Unacceptable	>20	>20	>20

Note. EMA = Ecological Momentary Assessment; SA = Suicide Attempt; NSSI = Non-Suicidal Self-injury; BPD = Borderline Personality Disorder; DERS = Difficulties with Emotion Regulation Scale; QIDS = Quick Inventory of Depressive Symptoms; RSQ = Response Style Questionnaire; BIS = Barrett Impulsiveness Scale.

improve model performance in this area.

4.1. Limitations and future directions

Our findings have several limitations. First, we examined SI, not SB. Additional research identifying the presence of SB is therefore needed. Second, our sample was comprised of predominantly white, psychiatrically hospitalized patients, which may limit study generalizability. Third, replication of findings using a novel dataset is needed to provide more robust validation of generalizability of our results. Fourth, while there is evidence suggesting no iatrogenic effects of repeated assessment of SI (Coppersmith et al., 2022a, 2022b; Law et al., 2015), it is possible that some reactivity to EMA items was experienced that could have impacted our findings. Fifth, research is needed to examine temporal ordering of effects. Sixth, we did not examine potential contributions of biological or neuroimaging factors in this study, and this is an important direction for future research. Seventh, determining precisely if or how RF models should be incorporated into clinical practice to aid in risk classification (including over simpler models) is outside the scope of the current study. However, while RF models may not have been

interpretable at the local (i.e., single prediction) level in the past, current methods allow for both global (Gini importance) *and* local (why this classification was made for this specific instance) explanations to facilitate model interpretation (Hatwell et al., 2020). Due to the novelty of machine learning approaches, additional research would be needed to provide guidance for effectively implementing such models in clinical practice with low provider burden. Eighth, our RF models had slightly higher false negative rates, which could be costly in real world applications; subsequent work should identify and explore ways to reduce the false negative rate in classification models.

Nevertheless, this study represents an important and novel contribution to the literature. It is the first to examine unique contributions of baseline and momentary risk factors in the classification of multiple characteristics of momentary SI. We found that both baseline and momentary features provide important information in both correctly classifying and differentiating individual characteristics of SI. Our results support the relevance of machine learning approaches for accurate identification of SI characteristics and underscore the importance of understanding the factors that differentiate and drive different characteristics of SI. Expansion of this work can support use of these models to guide intervention strategies.

CRediT authorship contribution statement

M.L. Bozzay: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **C.D. Hughes:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **C. Eickhoff:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis. **H. Schatten:** Writing – review & editing, Data curation. **M.F. Armey:** Writing – review & editing, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

None.

Acknowledgements

The authors would like to thank the research assistants in the CELL lab for their assistance in collecting the data involved in this manuscript.

Funding sources

This study was supported by grants from the National Institute of Mental Health (R01MH095786 and R01MH097741) to M. Armey.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2024.08.038.

References

Allen, K.J., Bozzay, M.L., Edenbaum, E.R., 2019. Neurocognition and suicide risk in adults. Curr. Behav. Neurosci. Rep. 1–15.
Armey, M.F., Brick, L., Schatten, H.T., Nugent, N.R., Miller, I.W., 2018. Ecologically assessed affect and suicidal ideation following psychiatric inpatient hospitalization. Gen. Hosp. Psychiatry 63, 89–96.
Beck, A.T., 1988. Beck Hopelessness Scale. The Psychological Corporation.
Bernstein, D.P., Fink, L., Handelsman, L., Foote, J., 1994. Initial reliability and validity of a new retrospective measure of child abuse and neglect. Am. J. Psychiatry 151 (8), 1132–1136.
Breiman, L., 2001. Random forests. Machine Learn. 45 (1), 5–32.
Bryan, C.J., Rudd, M.D., 2016. The importance of temporal dynamics in the transition from suicidal thought to behavior. Clin. Psychol. Sci. Pract. 23 (1), 21–25.
Bryan, C.J., Rozek, D.C., Butner, J., Rudd, M.D., 2019. Patterns of change in suicide ideation signal the recurrence of suicide attempts among high-risk psychiatric outpatients. Behav. Res. Ther. 120, 103392 https://doi.org/10.1016/j.brat.2019.04.001.
Caruana, R., Niculescu-Mizil, A., 2006. An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd International Conference on Machine Learning.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.
Colic, S., He, J.C., Richardson, J.D., Cyr, K.S., Reilly, J.P., Hasey, G.M., 2022. A machine learning approach to identification of self-harm and suicidal ideation among military and police veterans. J. Military Veteran Family Health 8 (1), 56–67.
Coppersmith, D.D., Fortgang, R.G., Kleiman, E.M., Millner, A.J., Yeager, A.L., Mair, P., Nock, M.K., 2022a. Effect of frequent assessment of suicidal thinking on its incidence and severity: high-resolution real-time monitoring study. Br. J. Psychiatry 220 (1), 41–43.
Coppersmith, D.D., Ryan, O., Fortgang, R., Millner, A., Kleiman, E., Nock, M., 2022b. Mapping the Timescale of Suicidal Thinking.
Czyz, E., Koo, H., Al-Dajani, N., King, C., Nahum-Shani, I., 2021a. Predicting short-term suicidal thoughts in adolescents using machine learning: developing decision tools to identify daily level risk after hospitalization. Psychol. Med. 1–10.
Czyz, E., Yap, J., King, C., Nahum-Shani, I., 2021b. Using intensive longitudinal data to identify early predictors of suicide-related outcomes in high-risk adolescents: practical and conceptual considerations. Assessment 28 (8), 1949–1959.
Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., Musacchio, K.M., Jaroszewski, A.C., Chang, B.P., Nock, M.K., 2017. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol. Bull. 143 (2), 187.
Galynker, I., Yaseen, Z.S., Cohen, A., Benhamou, O., Hawes, M., Briggs, J., 2017. Prediction of suicidal behavior in high risk psychiatric patients using an assessment of acute suicidal state: the suicide crisis inventory. Depress. Anxiety 34 (2), 147–158.
Gratz, K., Roemer, L., 2004. Multidimensional assessment of emotion regulation and dysregulation: development, factor structure, and initial validation of the difficulties in emotion regulation scale. J. Psychopathol. Behav. Assess. 26 (1), 41–54. https://doi.org/10.1023/B:JOBA.0000007455.08539.94.
Haglund, A., Lysell, H., Larsson, H., Lichtenstein, P., Runeson, B., 2019. Suicide immediately after discharge from psychiatric inpatient care: a cohort study of nearly 2.9 million discharges. J. Clin. Psychiatry 80 (2), 0.
Hatwell, J., Gaber, M.M., Azad, R.M.A., 2020. CHIRPS: explaining random forest classification. Artif. Intell. Rev. 53, 5747–5788.
Hufford, M.R., Shields, A.L., Shiffman, S., Paty, J., Balabanis, M., 2002. Reactivity to ecological momentary assessment: an example using undergraduate problem drinkers. Psychol. Addict. Behav. 16 (3), 205.
Jacobucci, R., Littlefield, A.K., Millner, A.J., Kleiman, E.M., Steinley, D., 2021. Evidence of inflated prediction performance: a commentary on machine learning and suicide research. Clin. Psychol. Sci. 9 (1), 129–134.
Just, M.A., Pan, L., Cherkassky, V.L., McMakin, D.L., Cha, C., Nock, M.K., Brent, D., 2023. Retraction Note: Machine Learning of Neural Representations of Suicide and Emotion Concepts Identifies Suicidal Youth. Nature Publishing Group UK London.
Kendall, P.C., Butcher, J.N., Holmbeck, G.N., 1999. Handbook of Research Methods in Clinical Psychology, 2nd ed. John Wiley and Sons.
Kleiman, E.M., Turner, B.J., Fedor, S., Beale, E.E., Huffman, J.C., Nock, M.K., 2017. Examination of real-time fluctuations in suicidal ideation and its risk factors: results from two ecological momentary assessment studies. J. Abnorm. Psychol. 126 (6), 726.
Law, M.K., Furr, R.M., Arnold, E.M., Mneimne, M., Jaquett, C., Fleeson, W., 2015. Does assessing suicidality frequently and repeatedly cause harm? A randomized control study. Psychol. Assess. 27 (4), 1171.
Menze, B.H., Kelm, B.M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., Hamprecht, F.A., 2009. A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinformatics 10 (1), 1–16.
Miller, I.W., Norman, W.H., Bishop, S.B., Dow, M.G., 1986. The modified scale for suicidal ideation: reliability and validity. J. Consult. Clin. Psychol. 54 (5), 724–725.
National Action Alliance for Suicide Prevention, R.T.F., 2014. A Prioritized Research Agenda for Suicide Prevention: An Action Plan to Save Lives. National Institute of Mental Health and the Research Prioritization Task Force, Rockville, MD.
Nolen-Hoeksema, S., Morrow, J., 1991. A prospective study of depression and posttraumatic stress symptoms after a natural disaster: the 1989 Loma Prieta earthquake. J. Pers. Soc. Psychol. 61 (1), 115–121. https://doi.org/10.1037/0022-3514.61.1.115 http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1991-33414-001&site=ehost-live.
Patton, J.H., Stanford, M.S., Barratt, E.S., 1995. Factor structure of the Barratt impulsiveness scale [Research Support, Non-U.S. Gov't]. J. Clin. Psychol. 51 (6), 768–774. http://www.ncbi.nlm.nih.gov/pubmed/8778124.
Posner, K., Brent, D., Lucas, C., Gould, M., Stanley, B., Brown, G., Fisher, P., Zelazny, J., Burke, A., Oquendo, M., 2008. Columbia-suicide Severity Rating Scale (C-SSRS). Columbia University Medical Center, New York, NY.
Rush, A.J., Trivedi, M.H., Ibrahim, H.M., Carmody, T.J., Arnow, B., Klein, D.N., Markowitz, J.C., Ninan, P.T., Kornstein, S., Manber, R., Thase, M.E., Kocsis, J.H., Keller, M.B., 2003. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol. Psychiatry 54 (5), 573–583.
Schafer, K.M., Kennedy, G., Gallyer, A., Resnik, P., 2021. A direct comparison of theory-driven and machine learning prediction of suicide: a meta-analysis. PloS One 16 (4), e0249833.
Selby, E.A., Joiner, T.E., 2013. Emotional cascades as prospective predictors of dysregulated behaviors in borderline personality disorder. Personal. Disord. Theory Res. Treat. 4 (2), 168.

Simons, J.S., Gaher, R.M., 2005. The Distress Tolerance Scale: development and validation of a self-report measure. Motiv. Emot. 29 (2), 83–102.

Van Orden, K.A., Witte, T.K., Gordon, K.H., Bender, T.W., Joiner Jr., T.E., 2008. Suicidal desire and the capability for suicide: tests of the interpersonal-psychological theory of suicidal behavior among adults. J. Consult. Clin. Psychol. 76 (1), 72–83. https://doi.org/10.1037/0022-006X.76.1.72.

Van Orden, K.A., Cukrowicz, K.C., Witte, T.K., Joiner Jr., T.E., 2012. Thwarted belongingness and perceived burdensomeness: construct validity and psychometric properties of the Interpersonal Needs Questionnaire. Psychol. Assess. 24 (1), 197.

Wang, S.B., Coppersmith, D.D., Kleiman, E.M., Bentley, K.H., Millner, A.J., Fortgang, R., Mair, P., Dempsey, W., Huffman, J.C., Nock, M.K., 2021. A pilot study using frequent inpatient assessments of suicidal thinking to predict short-term postdischarge suicidal behavior. JAMA Netw. Open 4 (3), e210591.

Watson, D., Clark, L., 1994. The PANAS-X: Manual for the Positive and Negative Affect Schedule – Expanded Form (Unpublished manuscript).

Weissman, A.N., Beck, A.T., 1978. Development and Validation of the Dysfunctional Attitude Scale: A Preliminary Investigation.

Zanarini, M.C., Vujanovic, A.A., Parachini, E.A., Boulanger, J.L., Frankenburg, F.R., Hennen, J., 2003. A screening measure for BPD: the Mclean Screening Instrument for Borderline Personality Disorder (MSI-BPD). J. Pers. Disord. 17 (6), 568–573. https://doi.org/10.1521/pedi.17.6.568.25355.