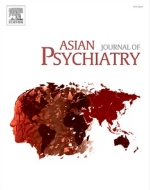
Contents lists available at ScienceDirect

Asian Journal of Psychiatry

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



An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach

Noratikah Nordin [a,*,1], Zurinahni Zainol [a,*,2], Mohd Halim Mohd Noor [a,*,3], Lai Fong Chan [b,4]

[a] School of Computer Sciences, Universiti Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia
[b] Department of Psychiatry, Faculty of Medicine, National University of Malaysia (UKM), 56000 Cheras, Wilayah Persekutuan Kuala Lumpur, Malaysia

ARTICLE INFO

Keywords:
Ensemble learning
SHAP
Explainable AI
Predictive model
Suicide attempt risk

ABSTRACT

Machine learning approaches have been used to develop suicide attempt predictive models recently and have been shown to have a good performance. However, those proposed models have difficulty interpreting and understanding why an individual has suicidal attempts. To overcome this issue, the identification of features such as risk factors in predicting suicide attempts is important for clinicians to make decisions. Therefore, the aim of this study is to propose an explainable predictive model to predict and analyse the importance of features for suicide attempts. This model can also provide explanations to improve the clinical understanding of suicide attempts. Two complex ensemble learning models, namely Random Forest and Gradient Boosting with an explanatory model (SHapley Additive exPlanations (SHAP)) have been constructed. The models are used for predictive interpretation and understanding of the importance of the features. The experiment shows that both models with SHAP are able to interpret and understand the nature of an individual's predictions with suicide attempts. However, compared with Random Forest, the results show that Gradient Boosting with SHAP achieves higher accuracy and the analyses found that history of suicide attempts, suicidal ideation, and ethnicity as the main predictors for suicide attempts.

1. Introduction

Suicide is a major public health problem and one of the leading causes of death worldwide. It is estimated that nearly one million people have died by suicide, and the number of suicide attempts has recently been estimated to be ten to twenty times higher (Franklin et al., 2017; O'Connor and Nock, 2014). Accurately predicting or identifying individuals at risk for future suicide attempts is a major challenge in psychiatry and particularly in patients with depression. Many studies have used conventional approaches to identify clinical risk factors that might help suicide risk in depression (Ahmed et al., 2017; Chan et al., 2011). However, these efforts have been mostly unsuccessful, and have resulted in many false positives or inconsistent findings across studies (Velupillai et al., 2019). The development of a multivariate or statistical framework that successfully incorporates the role of potential clinical and non-clinical risk factors could allow for accurate prediction of an individual's suicide attempt risk. If we are able to predict suicide risk at the individual level, we can undoubtedly improve efforts to reduce suicide attempts among high-risk patients in the general population (Boudreaux et al., 2021).

The causes of suicidal behaviour are complex and predicting suicide attempters and non-suicide attempters is a challenging classification problem involving the simultaneous presence of many potential risk factors such as psychological, biological, and environmental factors (Burke et al., 2019). In conjunction with recent advances in the field of artificial intelligence, there is increasing research on the application of machine learning to assist in the detection, prediction and treatment of suicidal behaviours, including suicidal ideations, suicidal attempts and self-harms (Dwyer et al., 2018). Several types of machine learning have been proposed to develop a predictive model for suicide attempts.

* Corresponding authors.
E-mail addresses: noratikahnordin@student.usm.my (N. Nordin), zuri@usm.my (Z. Zainol), halimnoor@usm.my (M.H. Mohd Noor), laifchan@ppukm.ukm.edu.my (L.F. Chan).
[1] *ORCID ID - 0000–0001-9259–4309*
[2] *ORCID ID - 0000–0001-8908–5263*
[3] *ORCID ID - 0000–0002-3300–3270*
[4] *ORCID ID - 0000-0003-0958-6178*

https://doi.org/10.1016/j.ajp.2022.103316
Received 12 August 2022; Received in revised form 20 October 2022; Accepted 4 November 2022
Available online 7 November 2022
1876-2018/© 2022 Elsevier B.V. All rights reserved.

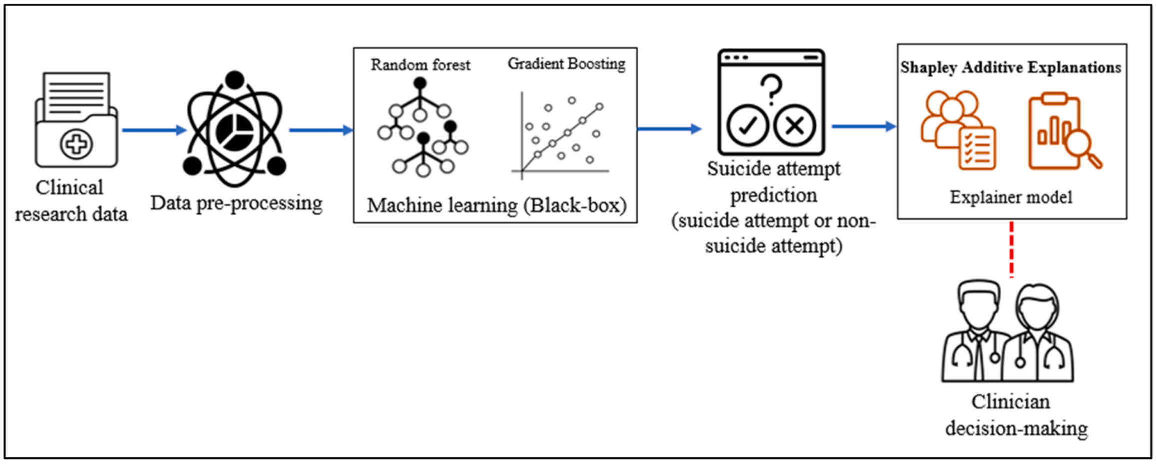

Clinical research data
Data pre-processing
Random forest
Gradient Boosting
Machine learning (Black-box)
Suicide attempt prediction (suicide attempt or non-suicide attempt)
Shapley Additive Explanations
Explainer model
Clinician decision-making

Fig. 1. Explainable predictive model for suicide attempt using Shapley Additive Explanations (SHAP).

Machine learning models are constructed to consider the complex relationships of risk factors to determine an ideal prediction model (Burke et al., 2020; Jung et al., 2019; Walsh et al., 2017). However, due to the complexity and dynamic characteristics of suicide attempts, it remains challenging to develop a universal predictive model that analyses and explains the risk factors for an individual. In fact, the complex machine learning models currently being developed are not able to provide a well-explained and interpretable prediction for decision efficiency (Kessler et al., 2020; Ryu et al., 2018).

The lack of interpretability becomes a challenging issue when the machine learning models are entrusted with the power to make clinical decisions that affect people's well-being (Knapič et al., 2021). To overcome this issue, explainable artificial intelligence (XAI) methods have been introduced to make the decision-making process of complex machine learning models more understandable to humans. The model-agnostic technique is one of the XAI methods specifically for model-related explainability, which is used on the machine learning model to provide post-hoc explanations (Abdullah et al., 2021). Although this model-agnostic is popular for feature-based XAI methods, there is still much to understand before the model can be adopted for clinical applicability. Therefore, the aim of this study is to propose an explainable predictive model for suicide attempt prediction to analyse feature importance and provide explanations to improve clinical understanding of suicide attempt risk prediction. We have focused on the prediction of suicide attempts using two ensemble machine learning models: random forest and gradient boosting, with Shapley Additive Explanations (SHAP Value), a model-agnostic method to analyse the importance of features that have a direct impact on the prediction of suicide attempt.

The paper is organized as follows: Section 2 presents the related works on the current approaches of suicide attempt prediction models. Section 3 discusses the proposed explainable predictive model and Section 4 presents the results and discussions of the proposed model, and finally, Section 5 summarizes the conclusion with future developments.

2. Related works

Research in clinical psychology and psychiatry has recently begun to use data mining and machine learning techniques to overcome the limitation of conventional statistical techniques (Bernert et al., 2020; Dwyer et al., 2018). Machine learning techniques are useful in classifying large numbers of patients into general risk categories and identifying potentially at-risk patients whose suicidality might otherwise have gone undetected (Fonseka et al., 2019). Edgcomb et al. (2021) proposed a classification and regression tree (CART) for predicting suicidal behaviour and self-injury in adults with serious mental illness. The results showed that CART is able to predict the risk of suicide attempt with good performance (accuracy – 0.80, AUC = 0.86, sensitivity = 0.79, specificity = 0.81). In a recent study by Kim et al. (2021), they compare random forest with k-Nearest Neighbours for detecting suicide risk in college students and the results show that random forest performs well (precision = 0.953).

Besides that, Navarro et al. (2021) proposed a model for suicide attempts in young people using random forest, while Cho et al. (2021) proposed the same technique for suicides in the elderly. Both studies highlighted the good performance of both random forest models in terms of specificity (0.76 – 0.833) and AUC (0.76 – 0.818). Ensemble prediction models such as boosting, bagging, random forest, and voting were found to perform better compared to a single prediction model such as decision tree, logistic regression and k-nearest neigbors (Nordin et al., 2021; Walsh et al., 2017). Although these studies are able to achieve good predictive performance, the contribution of the features to the predictions is not explained. Furthermore, due to the complexity of the ensemble models which are known to be black boxes that lack interpretability and explainability, it is almost impossible to understand the predictions (Boudreaux et al., 2021; Kessler et al., 2020; Ryu et al., 2018).

Explainable artificial intelligence (XAI) is defined as a method in the application of artificial intelligence such that the results of the prediction can be understood by human experts (Amann et al., 2020). According to Belle (Belle and Papantonis, 2021), explainability approaches are classified into transparent model and opaque model. Transparent machine learning models refer to models that can be easily interpreted and understandable by itself such as linear regression, logistic regression, decision tree, k-nearest neigbors and Bayesian model. Opaque machine learning models generally refers to the models that are difficult to interpret and understand but achieves higher accuracy such as random forest, artificial neural network, and support vector machine. In the opaque models, post-hoc explainability methods are introduced to understand how an already developed machine learning model produces its predictions for any given input (Barredo Arrieta et al., 2020). The post-hoc explainability methods can be categorized into model-specific and model-agnostic.

Model-specific refers to methods that are designed to explain and exploit the parameters based on their internal mechanisms such as structure or weights, and cannot readily be transferable to other models while the model-agnostic refers to methods that extract post-hoc explanations (explanations that are generated after the model has been trained) by treating the original model as a black box and not dependent on the structure of internal models (Belle and Papantonis, 2021). This study focused on the model-agnostic methods due to the model, explanation and representation flexibility (Ribeiro et al., 2016). Several

Table 1
Dataset description.

Data category	Data items (risk factors)	Values
Demographic	Gender	0 – Male, 1 – Female
	Ethnicity/Race	0 – Malay, 1 – Chinese, 2 - Indian
	Religion	0 – Muslim, 1 – Buddhist, 2 – Hindu, 3 - Christian
	Marital status	0 – Unmarried, 1 – Married
Clinical information	Psychotic features	0 – No, 1 – Yes
	Melancholic features	0 – No, 1 – Yes
	Suicidal ideation	0 – No, 1 – Yes
	Anxiety disorder	0 – No, 1 – Yes
	Severity of depression	0 – Mild, 1 – Moderate, 2 – Severe
	Medical problem	0 – No, 1 – Yes
	Nicotine dependence	0 – No, 1 – Yes
	Alcohol abuse	0 – No, 1 – Yes
	Any substance abuse	0 – No, 1 – Yes
	Sexual abuse	0 – No, 1 – Yes
	Mood stabilizer	0 – No, 1 – Yes
	History of hospitalization	0 – No, 1 – Yes
	Past suicide attempts	0 – No, 1 – Yes
	Family history of suicide attempts	0 – No, 1 – Yes

model-agnostic methods are Local Interpretable Model-Agnostic Explanation (LIME; (Ribeiro et al., 2016)), Anchors (Ribeiro et al., 2018) and Shapley Additive Explanation (SHAP; (Lundberg and Lee, 2017)). However, LIME and Anchors have drawbacks where the result of those methods are unstable generated explanations and sensitive to the dimensionality of the dataset (Sahakyan et al., 2021). This is because LIME and Anchors generates explanation based on random perturbations and when the number of features is increased, the local explanation is unable to discriminate the relevant features, which may provide poor performance and missing out on important features (Zafar and Khan, 2021).

Therefore, Shapley Additive exPlanation (SHAP) was introduced by Lundberg and Lee (2017) to overcome the problems. SHAP is a model-agnostic method that is based on game-theory inspired that attempts to enhance interpretability by computing the importance values for each feature for individual predictions. The SHAP calculates an additive feature importance score for each particular prediction that maintains three desirable properties: missingness, consistency and local accuracy (Linardatos et al., 2020). The SHAP is good at explaining and displaying how a feature value contributes to the prediction using SHAP values. The SHAP values provide a dynamic view of the effects of the interaction between the features to determine the probability of risk and the role of each feature on the individual level. In addition, the SHAP offers the possibility to visualize and explain the features responsible for the prediction at both local and global explanations (Abdullah et al., 2021; Belle and Papantonis, 2021). Therefore, the main contribution of this study is to propose an explainable model for predicting suicide attempts as well as provide clinicians with explanations of why a certain prediction is made and to analyse which risk factors lead to this prediction using the ensemble learning model and SHAP.

3. Method

Fig. 1 illustrates the proposed explainable predictive model for predicting suicide attempts. The proposed model provides explanations to improve the clinical understanding of risk prediction of suicide attempts. The clinical research data are obtained from a psychiatrist and the data are pre-processed to develop the predictive models. Two ensemble learning models (Random Forest and Gradient Boosting) are constructed, and then, the explanatory model (SHAP) is used to analyse the significance of the features and provide explanations of the predictions for clinicians' decision-making.

3.1. Data collection and data pre-processing

The study was conducted using clinical research data from the academic medical centre in Malaysia. The dataset consists of 75 psychiatric inpatients with depressive disorders (Chan et al., 2011). The dataset contains 18 variables including demographic, and clinical information about the patients as shown in Table 1. These features are used to train the proposed predictive model. There are no missing values in the dataset and the numerical data were normalized to a range of 0–1.

The features are assessed by clinicians using several instruments and measures. Patients were interviewed by clinicians using the Structured Clinical Interview for DSM-IV Axis-I Disorders, Clinical Version (SCID-I/CV). Substance use disorders and anxiety disorders were also assessed using the Axis-I diagnoses. In addition, the severity of depression was measured using the Beck Depression Inventory (BDI), while suicidal ideation was assessed using the Scale for Suicidal Ideation (SSI).

This study focuses on a specific group of patients with depressive disorders. It includes 75 people aged 18–76 years and consists of 33 males (44%) and 42 females (56%). The majority of patients were Chinese (40%), followed by Malays (28%) and Indians (17%). The patients were mainly married with 50 patients (66.7%) and 25 patients (33.3%) non-married (single, divorced, separated, widowed). From the descriptive analysis, 32 patients (42.7%) had attempted suicide in the past and 56 patients (74.7%) reported having suicidal ideation. In addition, 36 patients (48%) were found to have been hospitalized in the past.

3.2. Machine learning models

We build two ensemble learning models that are widely accepted in healthcare which are Random Forest and Gradient Boosting (Navarro et al., 2021; Walsh et al., 2017) for suicide attempt prediction. Random Forest is the state-of-the-art ensemble learning model, and it is an extension of bagging. The main difference is the incorporation of randomized feature selection (Zhou, 2012). When constructing a large number of decision trees, Random Forest first randomly selects a subset of features at each split selection step and then performs the usual split selection procedure within the selected subset of features (Alpaydin, 2010). Gradient boosting is a boosting model in which a strong model is built by combining weaker models in sequence. The collection of the weak models forms a robust classification model. Gradient boosting is a useful and powerful algorithm for building predictive models because it provides more accurate results and performs the optimization in function space, which makes the use of custom loss functions easier (Kantardzic, 2019).

3.3. Shapley Additive Explanations (SHAP)

We used machine learning-based feature importance methods to understand the importance of the features and provide an explanation of the complex models to the model's prediction. We want to know which risk factors (features) have the highest association with the outcome (suicide attempts). SHapley Additive exPlanations (SHAP) is a model-agnostic explanation that assigns an importance value to each input feature for a given prediction. SHAP is based on the principles of cooperative game theory and the importance value obtained by probabilistically calculating the contribution of players in the game to the final game outcome using the Shapley value (Lundberg and Lee, 2017). By formulating the features as players in a coalition game, the Shapley values can be calculated to learn to distribute the pay-out fairly. In this context, the players are the features that have been used in the predictive model. The interaction between features is considered a 'team' of features, with each feature being a member of the team responsible for driving the overall prediction. Therefore, the Shapley value is used and defined as the average marginal contribution of an instance of a feature among all possible coalitions (combinations) of features.

In the context of suicide attempt prediction, a set of risk factors or features age (f_1), gender (f_2), and depression (f_3) is known to classify an individual with suicide attempts. To distribute the classification of an individual fairly, it is intended to measure the contribution of each risk factor, that is, the Shapley value of every risk factor. To calculate the Shapley value of a given risk factor, the difference between the classification that is generated when the risk factor is present is calculated with respect to the classification that is generated when the risk factor is absent. The difference is known as the marginal contribution of the given risk factor to the current coalition. The calculation is done for each coalition (subgroup) that are generated where the risk factor that is able to classify an individual with suicide attempt is present. The mean of the differences (average marginal contribution) in all coalitions is obtained, which is known as the Shapley value.

The calculation of Shapley value in SHAP is shown in Eq. (1) where the features are used to calculate the ratio of the contribution of a specific feature based on the weight of the contribution of all features (Lundberg and Lee, 2017).

$$\varnothing_f(x) = \sum_{s \subseteq F \setminus f} \left[|s| \quad \times \quad \binom{|F|}{|s|} \right]^{-1} C \tag{1}$$

where ∅ denotes for shapley value of feature, f with x is the observation (prediction task), s is the subset of features, F is full set of features available (the number of elements of the original set) and C is the marginal contribution value of adding the feature, f to that subset which is calculated by Eq. (2):

$$C = [x(s \cup f) - x(s)] \tag{2}$$

where $x(s \cup f)$ is the subset that includes features in s with feature f, and $x(s)$ is the subset without feature, f.

SHAP values are based on the outcome of each possible coalition of features that should be considered to determine the importance of a single feature. The coalition is created based on the cardinality of a power set is 2^F. For example, if we have three features (f_1, f_2, f_3), the possible coalitions of features is $2^3 = 8$, which means there are 8 possible coalitions. Since x is the observation (prediction task) and we take f_2 feature as an example on how to compute the SHAP value. The marginal contribution by f_2 feature to the prediction model containing only f_2 as a feature which is shown in the following formula.

$$C_{f_2 \quad , \quad \{f_2 \quad \}}(x) = (x)(s \cup f_2 \quad) - (x)(s)$$

In order to obtain the overall effect of f_2 on the final prediction model, the marginal contribution of f_2 in all the possible coalitions, s where f_2 is presence are needed. All the marginal contributions of f_2 are then aggregated through a weighted average using the formula;

$$\begin{aligned} \varnothing_{f_2} \quad (x) = w_1 \quad & \times C_{f_2 \quad , \quad \{f_2 \quad \}}(x) + w_2 \\ & \times \quad C_{f_2 \quad , \quad \{f_2 \quad ,f_1 \quad \}}(x) + \quad w_3 \\ & \times \quad C_{f_2 \quad , \quad \{f_2 \quad ,f_3 \quad \}}(x) + \quad w_4 \quad \times \quad C_{f_2 \quad , \quad \{f_2 \quad ,f_1,f_3\}}(x) \end{aligned}$$

Where $w_1 + w_2 + w_3 + w_4 = 1$. All the weights of marginal contribution to f-feature should equal to each other, for each f which means that the sum of the weights of all the marginal contributions to one-feature-models is equal to the sum of the weights of all the marginal contributions to two-feature-models. Therefore, we calculate the SHAP value of f_2 for prediction of an individual with suicide attempt:

Table 2
Performance model for predictive modeling of suicide attempts.

Classification	Random Forest	Gradient Boosting
Accuracy	0.84	0.86
Precision	0.84	0.85
Specificity	0.82	0.83
Sensitivity	0.84	0.85
AUPRC (area under precision-recall curve)	0.83	0.84

$$\begin{aligned} \varnothing_{f_2} \quad (x) = \quad & \left[\left(1 \quad \times \quad \binom{3}{1} \right) \right]^{-1} \\ & \times \quad C_{f_2 \quad , \quad \{f_2 \quad \}}(x) + \left[\left(2 \quad \times \quad \binom{3}{2} \right) \right]^{-1} \\ & \times \quad C_{f_2 \quad , \quad \{f_2 \quad ,f_1 \quad \}}(x) + \left[\left(2 \quad \times \quad \binom{3}{2} \right) \right]^{-1} \\ & \times \quad C_{f_2 \quad , \quad \{f_2 \quad ,f_3 \quad \}}(x) + \left[\left(3 \quad \times \quad \binom{3}{3} \right) \right]^{-1} \\ & \times \quad C_{f_2 \quad , \quad \{f_2 \quad ,f_1,f_3\}}(x) \end{aligned}$$

The SHAP value of each feature for suicide attempt prediction are calculated based on the equation given to determine the contribution of each feature.

Three cross-validation was used to evaluate the predictive models to increase the generalization of the model and to avoid overfitting because the number of samples is limited (Nordin et al., 2021). The analysis was conducted entirely in the Python programming language (version 3.0) for the development of the proposed model. The performance of two ensemble learning models was evaluated based on the accuracy, specificity, sensitivity and AUPRC for predicting the classes (non-suicide attempter and suicide attempter).

4. Results and discussion

4.1. Model performance analysis

A dataset containing information of 75 patients is used to develop the proposed ensemble models using machine learning approaches. Table 2 shows the performance of each machine learning model in classifying suicide attempters and non-suicide attempters. Gradient boosting achieved the highest accuracy of 0.86, while random forest achieved a slightly lower accuracy of 0.84. The values for precision and specificity of the two ensemble predictive models ranged from 0.84 to 0.85 and 0.82–0.83, respectively. The sensitivity of the test indicates that the proportion of individuals who attempted suicide will have a positive result, which means that gradient boosting is good for identifying suicide attempters if the suicide attempters have a positive test (0.85). In addition, the area under precision-recall curve (AUPRC) also shows that gradient boosting is able to classify individuals with suicidal attempts with 0.84 compared to 0.83 for random forest.

4.2. Feature importance and interpretability of personalized suicide attempt risk prediction

The relative contribution of predictive factors to suicide attempt risk was assessed using a predictive model (random forest & gradient boosting) and the outcomes affecting the predictive model using the SHAP explainer were integrated. We calculate the mean SHAP values of random forest and gradient boosting to explain and compare the influence of the features. SHAP values are useful to show the contribution of each feature to an individual prediction (Peng et al., 2021; Ward et al., 2021). Fig. 2 and Fig. 3 show the feature impacts across all patients for the 18 features where each point indicates the impact of the feature on the samples.

The horizontal axis (x-axis) represents the SHAP value, which

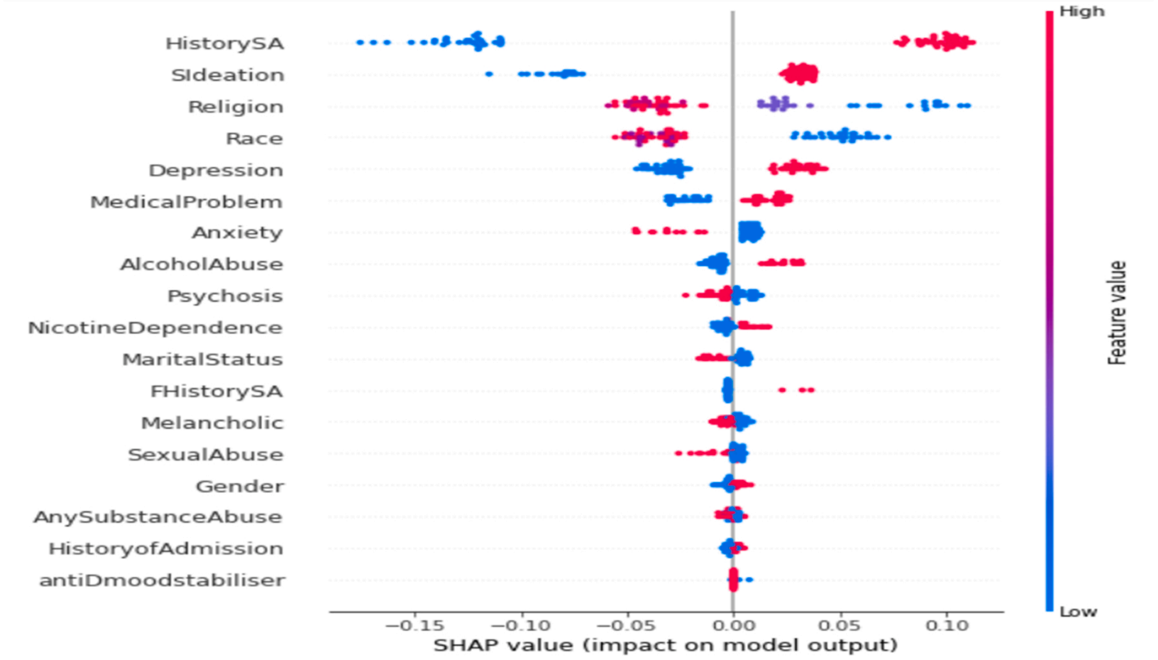


Fig. 2. Local explanation summary (averaged feature-importance) for the random forest.

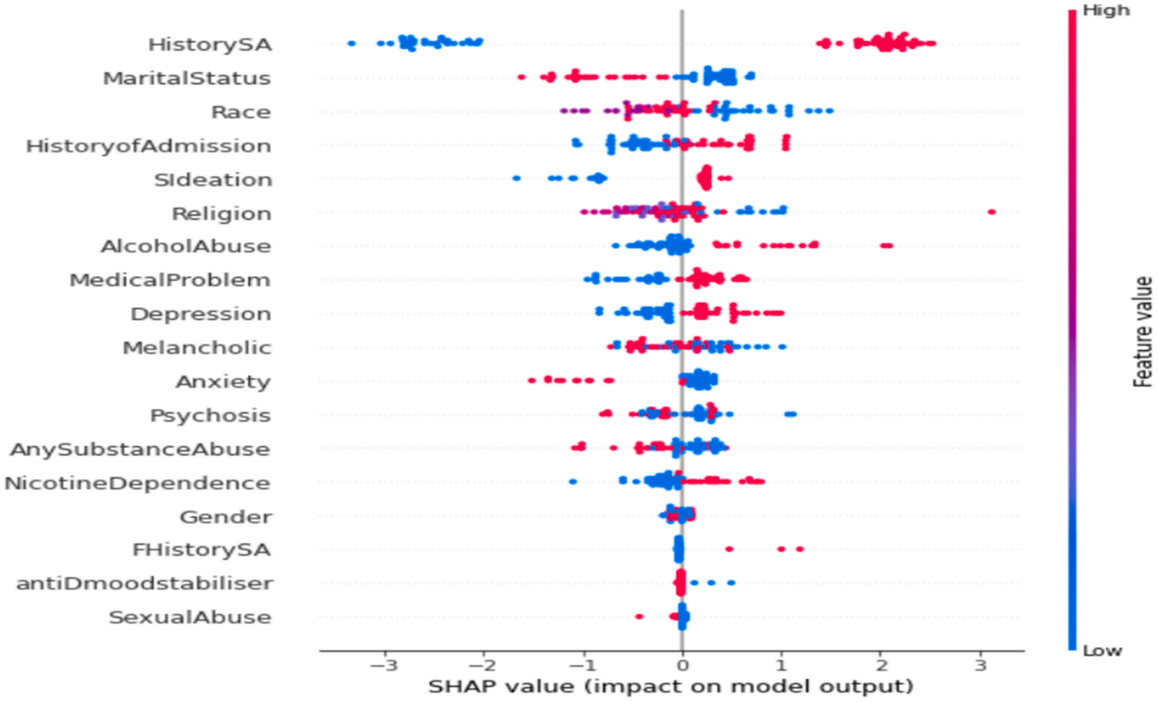


Fig. 3. Local explanation summary (averaged feature-importance) for gradient boosting.

denotes the average marginal contribution of the feature value to output across all possible coalitions. The SHAP value with less than 0 indicates a negative contribution, equal to 0 indicates no contribution, and greater than 0 indicates a positive contribution. Positive contribution means that the features have high importance to the final prediction, while the features that have the least important will lead to negative contribution. The features at the top contribute more to the model's prediction than the bottom, and we can see that each feature is ordered according to its importance as shown in Fig. 2. The 'history of suicide attempt' feature is the most important and 'mood stabilizer' feature is the least important. The longitudinal axis (y-axis) has two coordinates, left and right. The left longitudinal coordinate represents the features ranked by importance in descending order, while the right longitudinal coordinate indicates the value of the features from low to high. The color shows whether the feature is high (in red) or low (blue) for prediction. For Fig. 2 and Fig. 3, it can be seen that a high level of history of suicide attempts has a high and positive impact on the prediction of suicide attempt risk. The most important features that are positively correlated with suicide attempt risk for random forest are the history of suicide attempt, suicidal ideation, and severity of depression, as shown in Fig. 2.

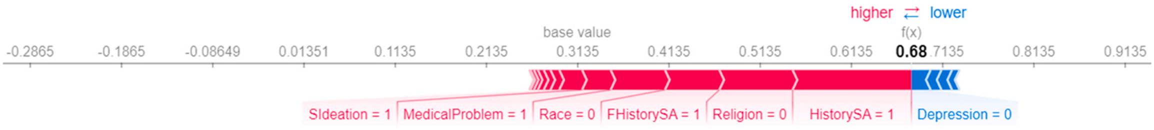


Fig. 4. Individual prediction of suicide attempt for patient 6.

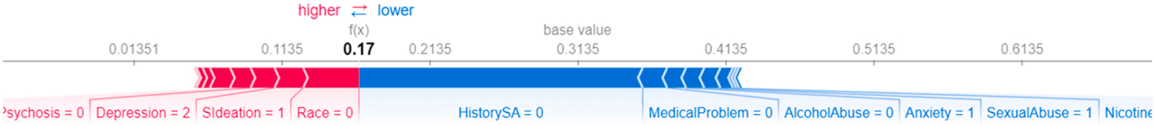


Fig. 5. Individual prediction of suicide attempt for patient 7.

In addition, the positively correlated and most important features for gradient boosting are history of suicide attempts, history of hospitalization and suicidal ideation, as shown in Fig. 3. Each of these features has an impact on prediction. The high feature value is indicated by the red colour, and the positive impact is indicated on the x-axis of the SHAP value.

To explain and interpret the prediction of suicide attempts in detail, we demonstrated and visualized an individual explanation of the model prediction using a force plot, as shown in Fig. 4 and Fig. 5. The force plot shows a prediction for two random patients, patient 6 and patient 7. The function *f(x)* is the output of the model (the predicted probability for this patient), and the base value follows the average of the model predictions. The features that increase (higher) the prediction are shown in red, while the features that decrease (lower) the prediction are shown in blue. In addition, the red features are right arrows, while the blue features are left arrows. The size of the arrow represents the effect of the features. In Fig. 4, we see that patient 6 has a high probability of suicide attempts (0.68) because of the risk factors that increase the prediction, such as has history of suicide attempts, has family history of suicide attempt, has medical problems, has suicidal ideation and mild depression.

In contrast to Fig. 4, patient 7 has a low probability of suicide attempts. This is because the risk factors (features) shift the prediction from the base value (0.31) to the model output (0.17). This means that the patient has no history of suicide attempts, no medical problems, no alcohol abuse, and no anxiety disorder, although the patient does have suicidal ideation.

Based on such individual explanations, we can make reliable decisions and provide clinicians with detailed information about which individuals are at high risk of attempting suicide in the future. With the given complex features and predictors, this will help clinicians in giving treating the patients. In addition, the feature importance by the SHAP model has been shown to improve the understanding of model performance compared to conventional machine learning models. The results of this study show that past suicide attempts, suicidal ideation, and race are the most important predictors of suicide attempts using random forest and gradient boosting. The predictor of suicidal ideation for suicide attempts is consistent with the findings of the studies by Chan et al. (2011) and Shen et al. (2020). Chan et al. (2011) found that risk factors for suicide attempts in depressed patients were suicidal ideation and alcohol use disorders, while the study by Ibrahim et al. (2017) showed that depression and anxiety were positively correlated with suicidal ideation.

Although suicidal ideation emerged as the most important predictor of suicide attempts, this study found that past suicide attempts and race were also important predictors of suicide attempts that were not discovered in previous studies (Mars et al., 2019; Walsh et al., 2018). For example, Mars et al. (2019) show that the strongest and most important predictors are suicidal thoughts and substance use disorders, and the Shen et al. (2020) study shows that suicide plans, anxiety, and depression were important contributors to suicide attempts. The ability of the SHAP model to explain the feature importance in local explanation is useful and helpful for medical decision-making.

Few studies have conducted predictive models for suicide attempts using machine learning approaches (Jung et al., 2019; Ryu et al., 2018; Walsh et al., 2017) and focused on accuracy in predicting an individual with suicidal behaviour. However, there are limitations to the prediction results using the corresponding predictive model, namely the reliability and transparency of the prediction result, which cannot be discovered due to an insufficient explanation of how the prediction was conducted for predicting suicide attempts. In this study, the SHAP value is proposed for the complex predictive model to predict the risk of suicide attempts by analysing the demographic and clinical factors.

5. Conclusion

In this study, we propose an explainable predictive model, that combines the complex models and the explanation model to reliably predict the risk of suicide attempts. A clinical dataset of patients with depression is used to evaluate the feasibility of the proposed model. The results show that gradient boosting achieves the best accuracy with 0.86, compared to random forest (0.84). The explanatory results generated by the proposed model can identify and explain the risk factors for suicide attempts and improve the understanding of suicide attempt prediction. The most important predictors contributing to the predictions of suicide attempts are individuals with past suicide attempts, and suicidal ideation. This study focuses on the value of explainable machine learning techniques in interpreting black-box models, which encourages the use of artificial intelligence in healthcare. Future research may focus on various machine learning models (decision tree, support vector machine, logistic regression) and investigate additional explainable machine learning techniques that can handle multicollinear features.

CRediT authorship contribution statement

NN and ZZ designed the study. NN and MHMN drafted the manuscript. LFC prepared the dataset and NN wrote the main manuscript text. ZZ, MHMN, and LFC critically revised the manuscript. All authors reviewed and approved the final manuscript.

Funding

This work has been supported by Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme (FRGS) with Project Code: FRGS/1/2020/ICT02/USM/02/5.

Conflict of Interest

The authors declared no conflicts of interest with respect to the authorship, research and publication of this article.

Acknowledgements

The authors would like to thank all parties that support and involve directly or indirectly into this research, especially Universiti Sains Malaysia, Pulau Pinang and Universiti Kebangsaan Malaysia Medical Centre, Kuala Lumpur.

References

Abdullah, T.A.A., Zahid, M.S.M., Ali, W., 2021. A review of interpretable ML in healthcare: taxonomy, applications, challenges, and future directions. Symmetry 13 (12), 2439. https://doi.org/10.3390/sym13122439.

Ahmed, H., Hossain, M., Aftab, A., Soron, T., Alam, M., Chowdhury, M.A., Uddin, A., 2017. Suicide and depression in the World Health Organization South-East Asia Region: A systematic review. WHO South-East Asia J. Public Health 6 (1), 60. https://doi.org/10.4103/2224-3151.206167.

Alpaydin, E., 2010. Introduction to Machine Learning, second ed. MIT Press.

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I., 2020. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inform. Decis. Mak. 20 (1), 310. https://doi.org/10.1186/s12911-020-01332-6.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.

Belle, V., Papantonis, I., 2021. Principles and practice of explainable machine learning. Front. Big Data 4, 688969. https://doi.org/10.3389/fdata.2021.688969.

Bernert, R.A., Hilberg, A.M., Melia, R., Kim, J.P., Shah, N.H., Abnousi, F., 2020. Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. Int. J. Environ. Res. Public Health 17 (16), 5929. https://doi.org/10.3390/ijerph17165929.

Boudreaux, E.D., Rundensteiner, E., Liu, F., Wang, B., Larkin, C., Agu, E., Ghosh, S., Semeter, J., Simon, G., Davis-Martin, R.E., 2021. Applying machine learning approaches to suicide prediction using healthcare data: overview and future directions. Front. Psychiatry 12, 707916. https://doi.org/10.3389/fpsyt.2021.707916.

Burke, T.A., Ammerman, B.A., Jacobucci, R., 2019. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: a systematic review. J. Affect. Disord. 245, 869–884. https://doi.org/10.1016/j.jad.2018.11.073.

Burke, T.A., Jacobucci, R., Ammerman, B.A., Alloy, L.B., Diamond, G., 2020. Using machine learning to classify suicide attempt history among youth in medical care settings. J. Affect. Disord. 268, 206–214. https://doi.org/10.1016/j.jad.2020.02.048.

Chan, L.F., Maniam, T., Shamsul, A.S., 2011. Suicide attempts among depressed inpatients with depressive disorder in a Malaysian sample: psychosocial and clinical risk factors. Crisis 32 (5), 283–287. https://doi.org/10.1027/0227-5910/a000088.

Cho, S.-E., Geem, Z.W., Na, K.-S., 2021. Development of a suicide prediction model for the elderly using health screening data. Int. J. Environ. Res. Public Health 18 (19), 10150. https://doi.org/10.3390/ijerph181910150.

Dwyer, D.B., Falkai, P., Koutsouleris, N., 2018. Machine learning approaches for clinical psychology and psychiatry. Annu. Rev. Clin. Psychol. 14 (1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Edgcomb, J.B., Shaddox, T., Hellemann, G., Brooks, J.O., 2021. Predicting suicidal behavior and self-harm after general hospitalization of adults with serious mental illness. J. Psychiatr. Res. 136, 515–521. https://doi.org/10.1016/j.jpsychires.2020.10.024.

Fonseka, T.M., Bhat, V., Kennedy, S.H., 2019. The utility of artificial intelligence in suicide risk prediction and the management of suicidal behaviors. Aust. N. Z. J. Psychiatry 53 (10), 954–964. https://doi.org/10.1177/0004867419864428.

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), 187–232. https://doi.org/10.1037/bul0000084.

Ibrahim, N., Amit, N., Che Din, N., Ong, H.C., 2017. Gender differences and psychological factors associated with suicidal ideation among youth in Malaysia. Psychol. Res. Behav. Manag. Volume 10, 129–135. https://doi.org/10.2147/PRBM.S125176.

Jung, J.S., Park, S.J., Kim, E.Y., Na, K.-S., Kim, Y.J., Kim, K.G., 2019. Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. PLOS ONE 14 (6), e0217639. https://doi.org/10.1371/journal.pone.0217639.

Kantardzic, M., 2019. DATA MINING: Concepts, models, methods, and algorithms. JOHN WILEY.

Kessler, R.C., Bossarte, R.M., Luedtke, A., Zaslavsky, A.M., Zubizarreta, J.R., 2020. Suicide prediction models: a critical review of recent research with recommendations for the way forward. Mol. Psychiatry 25 (1), 168–179. https://doi.org/10.1038/s41380-019-0531-0.

Kim, S., Lee, H.-K., Lee, K., 2021. Detecting suicidal risk using MMPI-2 based on machine learning algorithm. Sci. Rep. 11 (1), 15310. https://doi.org/10.1038/s41598-021-94839-5.

Knapič, S., Malhi, A., Saluja, R., Främling, K., 2021. Explainable artificial intelligence for human decision support system in the medical domain. Mach. Learn. Knowl. Extr. 3 (3), 740–770. https://doi.org/10.3390/make3030037.

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2020. Explainable AI: a review of machine learning interpretability methods. Entropy 23 (1), 18. https://doi.org/10.3390/e23010018.

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Artif. Itell. https://doi.org/10.48550/ARXIV.1705.07874.

Mars, B., Heron, J., Klonsky, E.D., Moran, P., O'Connor, R.C., Tilling, K., Wilkinson, P., Gunnell, D., 2019. Predictors of future suicide attempt among adolescents with suicidal thoughts or non-suicidal self-harm: a population-based birth cohort study. Lancet Psychiatry 6 (4), 327–337. https://doi.org/10.1016/S2215-0366(19)30030-6.

Navarro, M.C., Ouellet-Morin, I., Geoffroy, M.-C., Boivin, M., Tremblay, R.E., Côté, S.M., Orri, M., 2021. Machine learning assessment of early life factors predicting suicide attempt in adolescence or young adulthood. JAMA Netw. Open 4 (3), e211450. https://doi.org/10.1001/jamanetworkopen.2021.1450.

Nordin, N., Zainol, Z., Mohd Noor, M.H., Lai Fong, C., 2021. A comparative study of machine learning techniques for suicide attempts predictive model. Health Inform. J. 27 (1) https://doi.org/10.1177/1460458221989395.

O'Connor, R.C., Nock, M.K., 2014. The psychology of suicidal behaviour. Lancet Psychiatry 1 (1), 73–85. https://doi.org/10.1016/S2215-0366(14)70222-6.

Peng, J., Zou, K., Zhou, M., Teng, Y., Zhu, X., Zhang, F., Xu, J., 2021. An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. J. Med. Syst. 45 (5), 61. https://doi.org/10.1007/s10916-021-01736-5.

Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). Model-Agnostic Interpretability of Machine Learning. https://doi.org/10.48550/ARXIV.1606.05386.

Ribeiro, M.T., Sameer, S., Carlos, G., 2018. Anchors: high-precision model-agnostic explanations. Proc. AAAI Conf. Artif. Intell. 32 (1), 1–9.

Ryu, S., Lee, H., Lee, D.-K., Park, K., 2018. Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. Psychiatry Investig. 15 (11), 1030–1036. https://doi.org/10.30773/pi.2018.08.27.

Sahakyan, M., Aung, Z., Rahwan, T., 2021. Explainable artificial intelligence for tabular data: a survey. IEEE Access 9, 135392–135422. https://doi.org/10.1109/ACCESS.2021.3116481.

Shen, Y., Zhang, W., Chan, B.S.M., Zhang, Y., Meng, F., Kennon, E.A., Wu, H.E., Luo, X., Zhang, X., 2020. Detecting risk of suicide attempts among Chinese medical college students using a machine learning algorithm. J. Affect. Disord. 273, 18–23. https://doi.org/10.1016/j.jad.2020.04.057.

Velupillai, S., Hadlaczky, G., Baca-Garcia, E., Gorrell, G.M., Werbeloff, N., Nguyen, D., Patel, R., Leightley, D., Downs, J., Hotopf, M., Dutta, R., 2019. Risk assessment tools and data-driven approaches for predicting and preventing suicidal behavior. Front. Psychiatry 10, 36. https://doi.org/10.3389/fpsyt.2019.00036.

Walsh, C.G., Ribeiro, J.D., Franklin, J.C., 2017. Predicting risk of suicide attempts over time through machine learning. Clin. Psychol. Sci. 5 (3), 457–469. https://doi.org/10.1177/2167702617691560.

Walsh, C.G., Ribeiro, J.D., Franklin, J.C., 2018. Predicting suicide attempts in adolescents with longitudinal clinical data and machine learning. J. Child Psychol. Psychiatry 59 (12), 1261–1270. https://doi.org/10.1111/jcpp.12916.

Ward, I.R., Wang, L., Lu, J., Bennamoun, M., Dwivedi, G., Sanfilippo, F.M., 2021. Explainable artificial intelligence for pharmacovigilance: what features are important when predicting adverse outcomes. Comput. Methods Prog. Biomed. 212, 106415 https://doi.org/10.1016/j.cmpb.2021.106415.

Zafar, M.R., Khan, N., 2021. Deterministic local interpretable model-agnostic explanations for stable explainability. Mach. Learn. Knowl. Extr. 3 (3), 525–541. https://doi.org/10.3390/make3030027.

Zhou, Z.-H., 2012. Ensemble Methods: Foundations and Algorithms, first ed. Chapman and Hall/CRC. https://doi.org/10.1201/b12207.