Original Article

Healthc Inform Res. 2022 October;28(4):319-331.
https://doi.org/10.4258/hir.2022.28.4.319
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Discovery of Intentional Self-Harm Patterns from Suicide and Self-Harm Surveillance Reports

Vuttichai Vichianchai, Sumonta Kasemvilas

Hardware-Human Interface and Communications (H^2I-Comm) Laboratory, College of Computing, Khon Kaen University, Khon Kaen, Thailand

Objectives: The purpose of this study was to identify patterns of self-harm risk factors from suicide and self-harm surveillance reports in Thailand. **Methods:** This study analyzed data from suicide and self-harm surveillance reports submitted to Khon Kaen Rajanagarindra Psychiatric Hospital, Thailand. The process of identifying patterns of self-harm risk factors involved: data preprocessing (namely, data preparation and cleaning, missing data management using listwise deletion and expectation-maximization techniques, subgrouping factors, determining the target factors, and data correlation for learning); classifying the risk of self-harm (severe or mild) using 10-fold cross-validation with the support vector machine, random forest, multilayer perceptron, decision tree, k-nearest neighbors, and ensemble techniques; data filtering; identifying patterns of self-harm risk factors using 10-fold cross-validation with the classification and regression trees (CART) technique; and evaluating patterns of self-harm risk factors. **Results:** The random forest technique was most accurate for classifying the risk of self-harm, with specificity, sensitivity, and F-score of 92.84%, 93.12%, and 91.46%, respectively. The CART technique was able to identify 53 patterns of self-harm risk, consisting of 16 severe self-harm risk patterns and 37 mild self-harm risk patterns, with an accuracy of 92.85%. In addition, we discovered that the type of hospital was a new risk factor for severe self-harm. **Conclusions:** The procedure presented herein could identify patterns of risk factors from self-harm and assist psychiatrists in making decisions related to self-harm among patients visiting hospitals in Thailand.

Keywords: Data Adjustment, Machine Learning, Data Analysis, Self-Injurious Behavior, Suicide

Submitted: November 12, 2021
Revised: April 7, 2022
Accepted: September 7, 2022

Corresponding Author
Sumonta Kasemvilas
Hardware-Human Interface and Communications (H^2I-Comm) Laboratory, College of Computing, Khon Kaen University, 123 Moo 16 Mittraphap Road, Nai Mueang, Mueang Khon Kaen, Khon Kaen 40002, Thailand. Tel: +66 (0)9 4251 9245, E-mail: sumkas@kku.ac.th (https://orcid.org/0000-0002-7941-6150)

I. Introduction

According to the World Health Organization, nearly 800,000 people committed suicide in 2016, representing 1.4% of deaths worldwide [1]. Over three-quarters of these deaths occurred in low- to middle-income countries, and approximately 53,000 suicides were recorded in Thailand each year, representing the highest incidence in the Association of Southeast Asian Nations [2,3]. Due to the increasing number of people who are at risk for self-harm and commit suicide, it is vital to develop techniques for effectively monitoring patients who are at risk of self-harm.

Machine learning (ML), an application of artificial intelligence (AI), has been employed in healthcare information research to provide more rapid and accurate support for doctors' and psychiatrists' decisions [4]. In ML, a computer system is programmed with the ability to learn from its experiences and automatically improve. There are three main types of ML: supervised, unsupervised, and reinforcement learning. The supervised learning approach has been intensively used in healthcare informatics research [5,6].

With respect to mental wellness, ML techniques have been employed to classify suicide risk patterns. For example, Boonkwang et al. [7] classified suicidal ideation and patterns of suicidal ideation risk factors using the Iterative Dichotomiser 3 (ID3), J48, naive Bayes, and ensemble techniques. Zalar et al. [8] employed the decision tree (DT), genetic algorithm, and supplementary vector techniques to examine risk factors for attempting suicide. Edgcomb et al. [9] adapted the classification and regression trees (CART) technique to classify the risk of suicide attempts among women who had posthospital depression, bipolar disorder, and chronic psychosis. Although identifying self-harm risk patterns would be useful for psychiatrists, the main obstacle is obtaining accurate and suitable data for analysis.

Suicide and self-harm surveillance reports (RP.506S) are completed by hospitals in all 76 provinces in Thailand. The report form used for this dataset, which was designed by Khon Kaen Rajanagarindra Psychiatric Hospital with 10 versions from 2003 to the present, consists of the following data: type of hospital, region in Thailand, data source, demographic information, type of service, being hurt by others, hurt others, self-harm one or more times, self-harming methods, cause of depression, health problems, personal behavior, intervention, referring hospital type, thoughts of suicide, hospital admission, and death.

These data are reported to Khon Kaen Rajanagarindra Psychiatric Hospital, which is the center for collecting data for analysis by the Department of Mental Health in Thailand. Two types of services are provided in each visit: depression services and self-harm services. All visits were considered to have been made by patients at risk of self-harm because depressed patients are capable of self-harm, and self-harm patients are capable of repeating self-harm. Therefore, follow-up is necessary for both types of patients. Self-harm risk can be classified as severe and mild. A severe risk refers to self-harm that can lead to hospitalization or death, whereas a mild risk of self-harm is present in non-self-harm patients who may be followed up or first-time depression patients.

Therefore, the data from these surveillance reports are appropriate for determining the patterns of risk factors for self-harm. This study aimed to identify patterns of risk factors for self-harm using ML from RP.506S reports. The findings of this study will support psychiatrists' decision-making in the self-harm risk analysis of patients.

II. Methods

1. Dataset

The dataset used in this study was retrospectively collected from suicide and self-harm surveillance reports (RP.506S) from Khon Kaen Rajanagarindra Psychiatric Hospital recorded from 2004 to 2016. The reports refer to 192,234 visits from 103,316 patients: most patients had one visit (85,399 patients), followed by patients with two visits (12,089 patients), and patients with three visits (4,384 patients). One patient made 80 visits, and on average, each patient made two visits. There were 103 factors, including: (1) type of hospital (primary care, secondary care, tertiary hospital, psychiatric hospital, and other); (2) region in Thailand (northern, northeastern, central, eastern, western, and southern); (3) data source (i.e., the source of information provided to the patient at that visit, which was recorded by staff who interviewed the patient, asked close relatives or service providers, or made observations from the outpatient department card, inpatient medical records,-or death certificates or other sources); (4) the ethnicity of patients (Thai, hill tribes, Khmer, Myanmar, Laos, China, Western, and other); (5) demographic information (gender, age, marital status, and occupation); (6) type of service (e.g., depression or self-harm); (7) depression during the visit; (8) history of injury by abuse by others; (9) history of the patient harming another person; (10) the number of self-harm times (once and more than once); (11) self-harm methods; (12) problems that cause depression or self-harm; (13) health problems; (14) personal behavior; (15) intervention; (16) the type of hospital that referred the patient; (17) hospital admission; (18) death; and (19) suicidal ideation (if the patient did not die by suicide).

2. Conceptual Framework

Figure 1 shows the conceptual framework of this study, which can be described as follows: data preprocessing, classifying self-harm risk to select the dataset and the best ML method, data filtering, identifying patterns of self-harm risk factors, and evaluating the self-harm risk patterns.

1) Data preprocessing

This section describes the method of data preprocessing, including data preparation and cleaning, managing missing data, subgrouping factors, determining target classes, and data correlation.

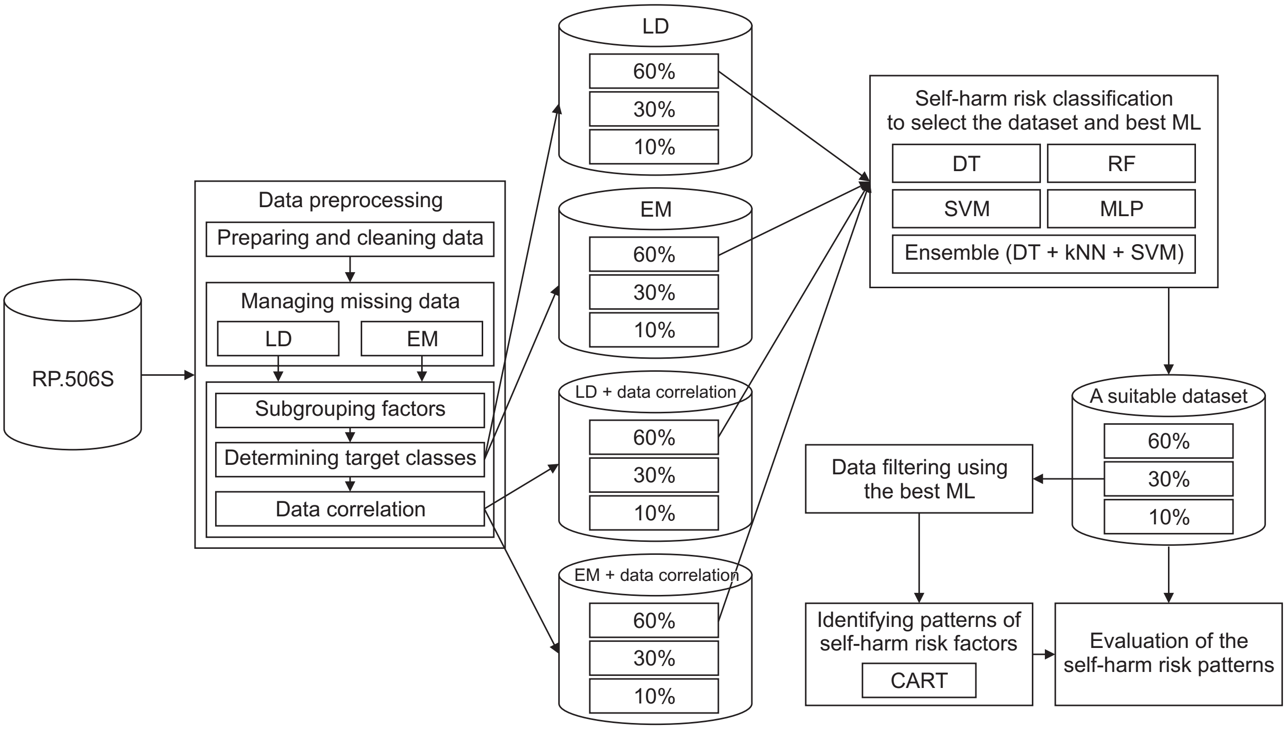

Figure 1. Conceptual framework of this research. LD: listwise deletion, EM: expectation-maximization, ML: machine learning, DT: decision tree, RF: random forest, SVM: support vector machine, MLP: multilayer perceptron, kNN: k-nearest neighbor, CART: classification and regression trees.

(1) Data preparation and cleaning

The following two steps were taken to increase the efficiency of classifying self-harm risk.

Step 1: Factors with multiple values were removed from the dataset.

Step 2: Factors with null values in more than 10% of all records were removed from the dataset.

(2) Missing data management

In this study, we chose two techniques for handling missing data: the listwise deletion (LD) technique and the expectation-maximization (EM) technique. The LD technique is a simple method of deleting records that contain missing data, but once the records are removed from the dataset, sufficient data must remain for analysis [10]. In most cases, if the number of records found to be lost does not exceed 10%–15%, the data can be deleted [10]. The EM technique is a complex method that can replace missing data without bias and is suitable for random distributions of the missing completely at random and missing at random types. It is a calculation based on maximum likelihood estimation using a parameter estimation method, consisting of two steps: expectation (E step), which involves the log-likelihood estimation of function parameters, and maximum value (maximization; M step), which replaces missing values with the values obtained from the E step. The expected values were re-estimated and compared until very little change was obtained, and that value was used to replace the missing data [11,12].

Step 1: Replace abnormal values with null values for all factors.

Step 2: Manage the missing data using the LD and EM techniques; each technique must be conducted separately.

(3) Subgrouping factors

Some factors needed to be categorized to optimize the classification of self-harm risk and to identify patterns of self-harm risk factors.

(4) Determine the target factors

This step included defining a target group for learning to identify self-harm risks and the patterns of self-harm risk factors. After targeting the datasets, in the next step, we randomly divided the data obtained from the LD and EM techniques into partitions of 60% (to classify self-harm risk), 30% (to find patterns of self-harm risk factor), and 10% (to assess the accuracy of the self-harm risk factor patterns).

(5) Data correlation

In this step, the correlation coefficients between variables in the dataset were calculated to eliminate variables that were strongly correlated. The correlation coefficient ranges between -1.0 and +1.0. A value near -1.0 indicates a strong negative correlation, a value near +1.0 indicates a strong positive correlation, and a value of 0 indicates that there is no correlation [13]. This technique reduces the number of variables in the dataset and solves the overfitting problem, which was important because this study used many variables; the use of many variables can lead to model overfitting, thereby decreasing the effectiveness of self-harm classification and the identification of self-harm risk factor patterns. After targeting the datasets, in the next step, we randomly divided the data obtained from the LD and EM techniques into partitions of 60% (to classify self-harm risk), 30% (to find patterns of the self-harm risk factor), and 10% (to assess the accuracy of the self-harm risk factor patterns).

2) Classifying the risk of self-harm

This step classified the risk for hospital admission from self-harm using popular ML techniques for suicidality risk classification to compare the effectiveness of techniques for managing missing data. We utilized the following techniques: support vector machine (SVM) [14-20], random forest (RF) [18,19,21-24], multilayer perceptron (MLP) [18-20], DT [7,9,18], and k-nearest neighbors (kNN) [25].

These steps were performed using a 2.5-GHz Dual-Core Intel Core i5 computer with 8 GB of RAM using Python version 2.7.16. The classification used 10-fold cross-validation with the DT, RF, SVM, MLP, and ensemble (voting for DT, kNN, and SVM) techniques.

3) Data filtering

This procedure filters the data using the best techniques obtained by classifying self-harm risk from 30% of an appropriate dataset. Data filtering compares the actual and predicted answers of the best techniques and selects all the correct prediction records.

4) Identifying the patterns of self-harm risk factors

This step was performed using a 2.5-GHz Dual-Core Intel Core i5 computer with 8 GB of RAM using WEKA version 3.8.5 to find the patterns of self-harm risk factors. The identification of patterns was performed using 10-fold cross-validation with the CART technique because it is simple to understand, assigns specific values to the inputs and outputs of each problem decision, and each probability can be evaluated [26].

5) Evaluation of the self-harm risk patterns

In this step, we compared the factors in each record (10% of an appropriate dataset) with the rules derived from the CART technique and calculated the accuracy rate based on the answers in the record.

3. Ethics

Ethics approval for using the dataset in this research was obtained from the Khon Kaen University Ethics Committee for Human Research (No. HE622093) and Ethics Committee on Human Research of Khon Kaen Rajanagarindra Psychiatric Hospital.

III. Results

1. Data Preprocessing

1) Data preparation and cleaning

Step 1: Eight factors with multiple values were removed, including the patient's name, patient’s family name, ID card number, patient identification number, visit dates, subdistrict codes, district codes, and addresses.

Step 2: Four factors with null values in more than 10% of all records were removed, including being hurt by others (19,880 visits or 10.34%), hurting others (177,125 visits or 92.14%), self-harm one or more times (110,494 visits or 57.48%) and having suicidal thoughts again (21,120 visits or 10.99%).

After data preparation and cleaning, 91 factors were left in the dataset for the next step.

2) Missing data management

The LD technique removes the records in which a null value is found. The factors in this dataset with null values are shown in Table 1. In this step, 6,601 records were deleted. The six factors for which null values most commonly caused record deletion were depression during the visit, referring hospital type, age, hospital admission, health problems, and death, with null values found in 5,293 records, 4,567 records, 3,479 records, 2,952 records, 2,551 records, and 1,425 records, respectively. Therefore, 185,633 records remained in the dataset after processing with the LD technique.

The EM technique was carried out using SPSS version 25 (IBM Corp., Armonk, NY, USA). The results for missing data management showed that the datasets managed using the LD and EM techniques had 185,633 and 192,234 visits, respectively.

3) Subgroupings of factors

In the study, two factors were categorized: (1) age was divided into five groups (i.e., under 18 years of age was represented by 1; 18–25 years of age was represented by 2; 26–45 years of age was represented by 3; 46–59 years of age was represented by 4, and older than 59 years of age was represented by 5; these age groups were classified based on the criteria of the Department of Mental Health, Ministry of Health, Thai-

Table 1. Demographics of patients and characteristics of their visits after data preprocessing

Factor	LD	EM
Types of hospital that is visited		
Primary care	5,909 (3.18)	6,324 (3.29)
Secondary care	92,457 (49.81)	94,673 (49.25)
Tertiary hospital	15,293 (8.24)	16,203 (8.43)
Psychiatric hospital	70,456 (37.95)	72,135 (37.52)
Other	1,518 (0.82)	2,899 (1.51)
Region of Thailand		
Northern	34,178 (18.41)	35,743 (18.59)
Northeastern	36,657 (19.75)	38,465 (20.01)
Central	45,988 (24.77)	46,289 (24.08)
Eastern	27,642 (14.89)	28,034 (14.58)
Western	18,034 (9.71)	18,945 (9.86)
Southern	24,134 (13.00)	24,758 (12.88)
Data source		
Patient	172,984 (93.19)	175,642 (91.37)
Relatives close people	54,491 (29.35)	56,703 (29.50)
Service provider	6,982 (3.76)	7,235 (3.76)
OPD card	72,657 (39.14)	73,768 (38.37)
Medical records - inpatients	19,456 (10.48)	21,008 (10.93)
Death certificate	67 (0.04)	89 (0.05)
Other	1,045 (0.56)	1,205 (0.63)
Ethnicity		
Thai	182,542 (98.33)	188,711 (98.17)
Hill tribe	1,729 (0.93)	1,794 (0.93)
Khmer	342 (0.18)	402 (0.21)
Myanmar	609 (0.33)	619 (0.32)
Laos	134 (0.07)	146 (0.08)
China	56 (0.03)	60 (0.03)
Western	77 (0.04)	80 (0.04)
Other	416 (0.22)	422 (0.22)
Gender		
Male	55,761 (30.04)	56,893 (29.60)
Female	129,872 (69.96)	135,341 (70.40)
Age (yr)		
<18	18,543 (9.99)	19,854 (10.33)
18–25	25,502 (13.74)	27,094 (14.09)
26–45	70,509 (37.98)	71,972 (37.44)
46–59	42,901 (23.11)	43,601 (22.68)
>59	28,178 (15.18)	29,713 (15.46)

Continued on the next page.

Table 1. Continued

Factor	LD	EM
Marital status		
Single	48,320 (26.03)	49,925 (25.97)
Married	108,728 (58.57)	110,965 (57.72)
Widowed	14,994 (8.08)	16,278 (8.47)
Divorced	10,746 (5.79)	11,846 (6.16)
Other	2,845 (1.53)	3,220 (1.68)
Occupation		
Agriculture	45,067 (24.28)	48,902 (25.44)
Employee/Labor	49,112 (26.46)	51,004 (26.53)
Housekeeper	22,008 (11.86)	22,789 (11.85)
Company worker	2,218 (1.19)	2,413 (1.26)
Trade/Personal business	13,099 (7.06)	14,289 (7.43)
Government officer/State enterprise	5,899 (3.18)	6,016 (3.13)
Student	14,533 (7.83)	15,067 (7.84)
Priest	1,356 (0.73)	1,542 (0.80)
Unemployed	6,123 (3.30)	6,420 (3.34)
Other	15,425 (8.31)	16,503 (8.58)
Type of service		
Service for depressed person		
First visit	86,721 (46.72)	88,372 (45.97)
Follow-up	63,984 (34.47)	64,209 (33.40)
Home visit	2,005 (1.08)	3,124 (1.63)
Depression during the visit		
No	53,712 (28.93)	57,124 (29.72)
Yes	131,921 (71.07)	135,110 (70.28)
Self-harm methods		
Overdose	31,234 (16.83)	33,876 (17.62)
Insecticide self-poisoning	7,207 (3.88)	7,312 (3.80)
Pesticide self-poisoning	6,342 (3.42)	6,488 (3.38)
Other chemicals self-poisoning	13,209 (7.12)	14,122 (7.35)
Injuring by sharp objects/solid substances	4,555 (2.45)	5,001 (2.60)
Firearm-related injury	378 (0.20)	402 (0.21)
Jumping from a height	607 (0.33)	712 (0.37)
Hanging	4,829 (2.60)	4,912 (2.56)
Drowning	406 (0.22)	512 (0.27)
Getting hit by a car	277 (0.15)	324 (0.17)
Poisoning by gas	5,912 (3.18)	6,088 (3.17)
Other	378 (0.20)	542 (0.28)

Continued on the next page.

Table 1. Continued

Factor	LD	EM
Problems or cause depression or self-harm		
To be offended by what other blame	29,230 (15.75)	29,786 (15.49)
To be scolded, i.e., chased to die, born again	1,134 (0.61)	1,280 (0.67)
Others gossip	504 (0.27)	609 (0.32)
To be disappointed in love	20,601 (11.10)	21,056 (10.95)
Quarrel with intimate partner	36,558 (19.69)	37,235 (19.37)
Want others to please me	6,541 (3.52)	6,893 (3.59)
Difficulties with learning	1,806 (0.97)	2,098 (1.09)
Lost family members	2,459 (1.32)	2,557 (1.33)
Quarrel among family members	3,704 (2.00)	2,836 (1.48)
Chronic disease/AIDS/disability	17,555 (9.46)	18,346 (9.54)
Afraid of HIV infection	450 (0.24)	626 (0.33)
Mental disorder	10,204 (5.50)	11,203 (5.83)
Depression	69,783 (37.59)	70,185 (36.51)
Elderly and living alone	1,955 (1.05)	2,109 (1.10)
Drug addiction (patient)	2,347 (1.26)	2,503 (1.30)
Drug addiction (family members)	2,288 (1.23)	2,356 (1.23)
Alcohol addiction	5,320 (2.87)	5,344 (2.78)
Poverty	11,774 (6.34)	12,135 (6.31)
Losing in business/Bankrupt	1,734 (0.93)	1,867 (0.97)
Loss property/Accident/Loss gambling	1,379 (0.74)	1,423 (0.74)
Lawsuit/Escape offense	702 (0.38)	812 (0.42)
Work problems	3,278 (1.77)	3,317 (1.73)
Unemployed	1,459 (0.79)	1,503 (0.78)
Other issues	12,534 (6.75)	12,760 (6.64)
Unknown	1,056 (0.57)	1,239 (0.64)
Health problems		
No health problems	104,934 (56.53)	106,485 (55.39)
Diabetes	13,566 (7.31)	13,572 (7.06)
Hypertension	21,489 (11.58)	21,490 (11.18)
Heart disease	3,444 (1.86)	3,498 (1.82)
Chronic kidney failure	1,203 (0.65)	1,288 (0.67)
Chronic liver disease	604 (0.33)	656 (0.34)
Chronic lung disease	1,588 (0.86)	1,689 (0.88)
Chronic headache	3,422 (1.84)	3,467 (1.80)
Arthritis/Gout/Back pain	5,743 (3.09)	5,790 (3.01)
Neuralgia	2,723 (1.47)	2,787 (1.45)
Epilepsy	1,975 (1.06)	2,023 (1.05)
Disabled	902 (0.49)	912 (0.47)
Paralysis/Stroke/Spinal cord	1,677 (0.90)	1,724 (0.90)
Cancer	932 (0.50)	945 (0.49)
Psychosis	9,567 (5.15)	9,583 (4.99)
AIDS/HIV	1,968 (1.06)	2,056 (1.07)
Other	15,251 (8.22)	15,884 (8.26)

Continued on the next page.

Table 1. Continued

Factor	LD	EM
Personal behavior		
Smoking addiction	11,877 (6.40)	12,008 (6.25)
Alcoholic	12,454 (6.71)	13,895 (7.23)
Drug addiction	3,599 (1.94)	4,004 (2.08)
Gambling addiction	818 (0.44)	980 (0.51)
Games/Internet addiction	666 (0.36)	712 (0.37)
Other	128,673 (69.32)	129,235 (67.23)
Intervention		
Psychotherapy consultation	143,688 (77.40)	147,243 (76.60)
Recommendation to patient's relatives	56,653 (30.52)	57,123 (29.72)
Knowledge documents	20,450 (11.02)	21,056 (10.95)
Amitriptyline or nortriptyline	36,177 (19.49)	38,239 (19.89)
Fluoxetine or SSRI	59,029 (31.80)	59,873 (31.15)
Other antidepressants	14,777 (7.96)	15,203 (7.91)
Anxiolytics	61,411 (33.08)	62,106 (32.31)
Psychotic drugs	21,783 (11.73)	22,290 (11.60)
Electric shock	169 (0.09)	224 (0.12)
Give money or stuff to patient	81 (0.04)	102 (0.05)
Enter the self-help group	719 (0.39)	902 (0.47)
Social service	745 (0.40)	1,004 (0.52)
Other	10,320 (5.56)	11,508 (5.99)
Types of hospital that is referred		
No refer	6,429 (3.46)	6,856 (3.57)
Primary care	148,367 (79.92)	151,765 (78.95)
Secondary care	19,114 (10.30)	20,567 (10.70)
Tertiary hospital	6,996 (3.77)	7,340 (3.82)
Psychiatric hospital	2,295 (1.24)	2,724 (1.42)
Other	2,432 (1.31)	2,982 (1.55)
Target classes		
Mild self-harm risk	100,758 (54.28)	105,359 (54.81)
Severe self-harm risk	84,875 (45.72)	86,875 (45.19)

Values are presented as number of visits (%).

LD: listwise deletion, EM: expectation-maximization, OPD: outpatient department, AIDS: acquired immune deficiency syndrome, HIV: human immunodeficiency virus, SSRI: selective serotonin reuptake inhibitor.

land [27]) and (2) the region of Thailand (northern, northeastern, central, eastern, western, and southern).

4) Determination of the target factors

There are two target classes for this study: severe self-harm risk and mild self-harm risk. Severe self-harm risk was determined based on hospital admission, death, or self-harm. Mild self-harm risk, which involved no hospitalization, death, or self-harm, is presented in Table 1.

After defining the target classes, we randomly divided the data into two datasets as follows: (1) The dataset managed using the LD technique was randomly divided into 111,380 visits, 55,690 visits, and 18,563 visits to classify self-harm risk, filter the data to find patterns in self-harm risk factors, and evaluate the patterns of self-harm risk factors, respectively. (2) The dataset managed using the EM technique was randomly divided into 115,340 visits, 57,670 visits, and 19,224 visits, which were used to classify self-harm risk, filter the data to find patterns in self-harm risk factors, and evaluate the patterns of self-harm risk factors, respectively.

5) Data correlations

In the study, factors with correlation coefficients of less than -0.9 or greater than +0.9 were selected from the dataset in Table 1. Sixty-two factors were found in the LD-processed dataset with correlation coefficients less than -0.9 or greater than +0.9 (29 factors were not deleted). In the EM-processed dataset, 57 factors had correlation coefficients less than -0.9 or greater than +0.9 (34 factors were not deleted).

After examining the correlations, we randomly divided the two datasets as follows: (1) The dataset managed using the LD technique and processed using data correlation was randomly divided into 111,380 visits, 55,690 visits, and 18,563 visits to classify self-harm risk, filter data to find patterns in self-harm risk factors, and evaluate the patterns of self-harm risk factors, respectively. (2) The dataset managed using the EM technique and processed using data correlation was randomly divided into 115,340 visits, 57,670 visits, and 19,224 visits to classify self-harm risk, filter data find patterns in self-harm risk factors, and evaluate the patterns of self-harm risk factors, respectively.

2. Classifying the Risk of Self-Harm

Based on the results presented in Tables 2 and 3, the LD technique and data correlation with the RF technique were most accurate for classifying the risk of self-harm, with area under the curve, specificity, sensitivity, and F-score of 97.52%, 92.84%, 93.12%, and 91.46%, respectively, based on 10-fold cross-validation.

3. Data Filtering

This procedure filtered data using the RF technique from 55,690 visits from the dataset processed using the LD and

Table 2. Classification of the risk of self-harm using ML techniques from the dataset using the LD and EM techniques

ML	Dataset using LD technique					Dataset using EM technique				
	AUC (%)	F-score (%)	SP (%)	SN (%)	Time (s)	AUC (%)	F-score (%)	SP (%)	SN (%)	Time (s)
Decision tree	95.21	88.56	89.98	88.45	17.56	94.01	87.13	87.40	86.98	22.12
Support vector machine	96.67	88.96	89.98	88.86	15,371.57	95.88	88.84	89.73	88.38	26,996.27
Random forest	96.62	90.12	90.59	89.58	142.02	96.04	89.15	90.59	88.50	156.61
Multilayer perceptron	89.51	88.68	88.91	88.58	9,983.65	88.21	88.19	88.04	88.21	11,123.43
Ensemble	88.65	87.88	89.12	86.58	16,120.96	87.40	87.30	88.81	85.85	33,144.49

ML: machine learning, LD: listwise deletion, EM: expectation-maximization, AUC: area under the curve, SP: specificity, SN: sensitivity.

Table 3. Classification of the risk of self-harm using ML methods from the datasets processed with the LD and EM techniques and data correlation

ML	Dataset using LD technique with data correlation					Dataset using EM technique with data correlation				
	AUC (%)	F-score (%)	SP (%)	SN (%)	Time (s)	AUC (%)	F-score (%)	SP (%)	SN (%)	Time (s)
Decision tree	95.21	88.56	90.00	88.45	8.78	94.01	87.13	87.42	86.98	11.06
Support vector machine	97.02	89.89	90.65	89.85	7,685.79	96.23	89.77	90.40	89.37	13,498.14
Random forest	97.52	92.84	93.12	91.46	71.01	96.94	91.87	93.12	90.38	78.21
Multilayer perceptron	90.89	88.69	88.92	88.68	5,471.89	89.59	88.20	88.05	88.31	5,689.55
Ensemble	88.66	87.89	89.13	86.59	9,163.23	87.41	87.31	88.82	85.86	17,523.35

ML: machine learning, LD: listwise deletion, EM: expectation-maximization, AUC: area under the curve, SP: specificity, SN: sensitivity.

Table 4. 16 patterns of severe self-harm risk factors

Pattern number	Pattern of severe self-harm risk factors
1	Depressed during visits → visiting a primary care → lived in the North or Northeast or the East or the South of the country
2	Depressed during visits → had depression problem → not visiting a secondary or a tertiary hospital
3	Depressed during visits → not visiting a primary care → got information from patient
4	Depressed during visits → not visiting a primary care → no information from the patient → lived in the Eastern or the South of the country
5	Depressed during visits → not visiting a primary care → no information from the patient → not living in the Eastern and the South of the country → quarrel with intimate partner
6	Depressed during visits → not visiting a primary care → no information from the patient → not living in the Eastern and the South of the country → intervention by referring the patient's relatives
7	Depressed during visits → visiting a primary care → no information from the patient → not living in the Eastern and the South of the country → under 45 years old → overdose
8	Depressed during visits → visiting a tertiary hospital → no information from the patient → not living in the Eastern or the South of the country → and under 45 years old
9	Depressed during visits → not visiting a primary care → no information from the patient → not living in the Eastern or the South of the country → over 45 years old → eat other chemicals self-poisoning
10	Depressed during visits → not visiting a primary care → no information from the patient → lived in the Northern or Northeastern of the country → over 45 years old
11	No depressed during visits → eat insecticide self-poisoning
12	No depressed during visits → self-harm → visiting a secondary care or a tertiary hospital or the other → got data source from patient
13	No depressed during visits → self-harm → visiting a secondary care or a tertiary hospital or the other → got data source from relative patient
14	No depressed during visits → self-harm → visiting a secondary care or a tertiary hospital or the other → intervention by referring the patient's relatives
15	No depressed during visits → self-harm → visiting a secondary care or a tertiary hospital or the other → drug addiction
16	No depressed during visits → self-harm → visiting a tertiary hospital or the other → having a career in agriculture or an employee or a laborer or a priest

data correlation techniques. The RF technique correctly predicted 51,235 visits and incorrectly predicted 4,455 visits, corresponding to an accuracy and inaccuracy of 92% and 8%, respectively.

4. Patterns of Self-Harm Risk Factors

There were 53 patterns of self-harm risk using the CART technique based on 10-fold cross-validation, including 16 patterns of severe self-harm risk factors, as shown in Table 4.

Table 4 shows patterns of severe self-harm risk, presenting the meaning of each pattern for example, the ninth pattern has the following meaning: if depressive symptoms were present during the visit and the patient visited primary care, information was not received from the patient, the patient did not live in the eastern or southern region of the country, the patient was over 45 years of age, and the patient had self-harm (self-poisoning by consuming other chemicals), it was concluded that the patient was at severe risk of self-harm. As another example, the 11th pattern means that if depressive symptoms were not present during the visit and the patient had self-harm (self-poisoning by eating insecticide), it was concluded that the patient was at severe risk of self-harm. These patterns are illustrated in the decision tree in Figure 2.

5. Evaluation of the Self-Harm Risk Patterns

The results of the self-harm risk pattern evaluation using 18,563 visits from the appropriate datasets found that 53 self-harm risk patterns were able to accurately identify 17,235 visits of self-harm risk, with an accuracy and inaccuracy of 92.85% and 7.15%, respectively.

https://doi.org/10.4258/hir.2022.28.4.319

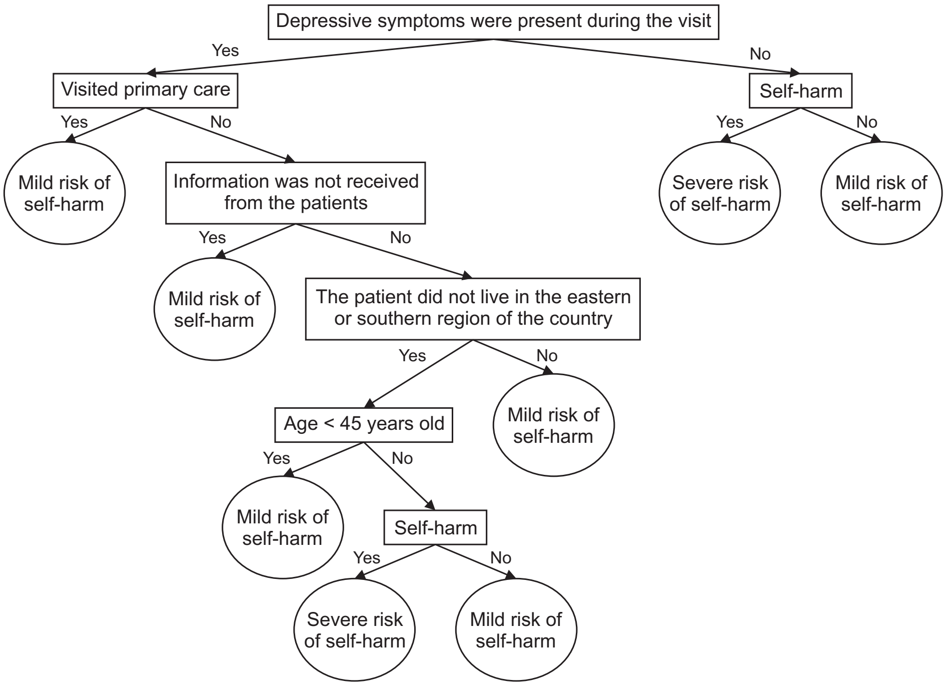

Figure 2. Example of a decision tree from the CART technique. CART: classification and regression trees.

IV. Discussion

This study proved that ML techniques combined with the LD method for handling missing data provided more accurate results and required less time than the EM technique. This was because the EM technique did not replace the missing data correctly, and because the EM technique is not biased when lost data have a random loss distribution [11,12], but this dataset had a nonrandom missing distribution factor, thereby causing bias. There was likely to be a nonrandom distribution of null values for personal behavior (i.e., smoking addiction, alcohol consumption, drug addiction, gambling addiction, and others) because Thai women are less likely to provide their personal behavior information than men, especially when the information reflects their bad habits. For example, Thai women patients with drug addiction and alcohol consumption may be likely to provide null values for these behaviors if they do not want to disclose.

In this study, we used many factors, which could lead to the overfitting problem. Therefore, this problem was solved by examining correlations to reduce the number of factors. Reducing the number of factors increased the efficiency of the self-harm risk classification, consistent with previous studies [13].

To classify the risk of self-harm, we chose popular ML techniques that have been applied to suicide risk classification in previous studies. The results of the self-harm risk classification using the SVM, RF, MLP, DT, and kNN techniques were satisfactory, with an accuracy of more than 87%, proving that these techniques are suitable for use in the classification of self-harm risks [14-25]. The RF technique is most effective because it reduces overfitting in decision trees and improves accuracy and flexibility in problem classification, which corresponds to high accuracy according to prior research [17-19,21-24].

In addition, we identified 53 patterns of self-harm risk factors using the CART technique, which was highly accurate because the CART technique assigns specific values to the inputs and outputs of each decision and can work well with correlated data [26]. Most of the risk factors found in this study were consistent with previous research, including depression, agricultural workers or laborers, being under 27 years of age, region of residence, and self-harm methods [7,9]. Two important factors appeared in patterns of severe self-harm risk factors: depression and the type of hospital. People with depression feel bored, angry, and worthless, causing them to have suicidal thoughts and putting them at risk for suicide. This may lead to suicide reattempts, as found in previous studies among Japanese people [28] and Thai policemen [29]. Regarding the type of hospital, patients with minor injuries and psychiatric symptoms were generally brought to a psychiatric hospital or a specialist clinic, patients with minor injuries and no psychiatric symptoms were brought to primary care, and patients with severe injuries were admitted to a tertiary hospital to treat wounds or symptoms. Therefore, we would like to suggest that all levels of hospitals should provide specialist psychiatrists to support mental treatment after the physical injury has healed. The findings of this study extend our knowledge of risk factors and identify the type of hospital as another risk factor. These findings may encourage other researchers to consider the type of hospital when examining factors influencing self-harm.

The RP.506S dataset, which contains patient self-harm surveillance data collected from hospitals in all 76 provinces of Thailand, has the advantage of being representative of the Thai population. However, the limitations of this study were as follows: (1) we were unable to manipulate null and invalid data in the dataset, and (2) the patterns of self-harm risk factors in this study may be inaccurate if applied in other countries because the data collected were designed for use with hospitals in Thailand.

In future studies, the proposed framework can be extended to a multiclass classification problem that could include the classification of depression severity among self-harm patients (normal, mild, moderate, and severe) or a classification of types of self-harm (escape, suicide, violence, and accident). Although identifying patterns of self-harm risk may not be sufficient to diagnose self-harm patients, it may assist psychiatrists in making decisions regarding hospital admission for self-harm patients.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

We would like to express our gratitude to the Director and staff of Khon Kaen Rajanagarindra Psychiatric Hospital for their support with data collection. This research is part of the dissertation titled "An algorithm for suicide prediction using machine learning techniques" by Mr. Vuttichai Vichianchai, PhD student, Khon Kaen University.

ORCID

Vuttichai Vichianchai (https://orcid.org/0000-0002-1222-1092)
Sumonta Kasemvilas (https://orcid.org/0000-0002-7941-6150)

References

1. World Health Organization. Mental health and substance use [Internet]. Geneva, Switzerland: World health Organication; 2020 [cited at 2022 Oct 26]. Available from: https://www.who.int/teams/mental-health-and-substance-use/data-research/suicide-data.
2. National News Bureau of Thailand. Warning issued over suicide rates among workers and unemployed [Internet]. 2022 [cited at 2022 Oct 27]. Available from: https://thainews.prd.go.th/en/news/detail/TCATG220503095140498.
3. The ASEAN Post. Suicides on the rise in Thailand [Internet]. Kuala Lumpur, Malaysia: The ASEAN Post; 2019 [cited at 2022 Oct 26]. Available from: https://theaseanpost.com/article/suicides-rise-thailand.
4. Mansourian M, Khademi S, Marateb HR. A comprehensive review of computer-aided diagnosis of major mental and neurological disorders and suicide: a biostatistical perspective on data mining. Diagnostics (Basel) 2021;11(3):393. https://doi.org/10.3390/diagnostics11030393
5. Mitchell TM. Does machine learning really work? AI Mag 1997;18(3):11. https://doi.org/10.1609/aimag.v18i3.1303
6. Tougui I, Jilbab A, Mhamdi JE. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. Healthc Inform Res 2021;27(3):189-99. https://doi.org/10.4258/hir.2021.27.3.189
7. Boonkwang K, Kasemvilas S, Kaewhao S, Youdkang O. A comparison of data mining techniques for suicide attempt characteristics mapping and prediction. Proceedings of 2018 International Seminar on Application for Technology of Information and Communication; 2018 Sep 21-22; Semarang, Indonesia, p. 488-93. https://doi.org/10.1109/ISEMANTIC.2018.8549835
8. Zalar B, Kores Plesnicar B, Zalar I, Mertik M. Suicide and suicide attempt descriptors by multimethod approach. Psychiatr Danub 2018;30(3):317-22. https://doi.org/10.24869/psyd.2018.317
9. Edgcomb JB, Thiruvalluru R, Pathak J, Brooks JO. Machine learning to differentiate risk of suicide attempt and self-harm after general medical hospitalization of women with mental illness. Med Care 2021;59:S58-S64. https://doi.org/10.1097/mlr.0000000000001467
10. Myers TA. Goodbye, listwise deletion: presenting hot deck imputation as an easy and effective tool for handling missing data. Commun Methods Meas 2011;5(4):297-310. https://doi.org/10.1080/19312458.2011.624490
11. Ibrahim JG, Zhu H, Tang N. Model selection criteria for missing-data problems using the EM algorithm. J Am Stat Assoc 2008;103(484):1648-58. https://doi.org/10.1198/016214508000001057

12. Little TD, Jorgensen TD, Lang KM, Moore EW. On the joys of missing data. J Pediatr Psychol 2014;39(2):151-62. https://doi.org/10.1093/jpepsy/jst048
13. Subramanian J, Simon R. Overfitting in prediction models: is it a problem only in high dimensions? Contemp Clin Trials 2013;36(2):636-41. https://doi.org/10.1016/j.cct.2013.06.011
14. Lee SY, Lu RB, Wang LJ, Chang CH, Lu T, Wang TY, et al. Serum miRNA as a possible biomarker in the diagnosis of bipolar II disorder. Sci Rep 2020;10(1):1131. https://doi.org/10.1038/s41598-020-58195-0
15. Wang Y, Sun K, Liu Z, Chen G, Jia Y, Zhong S, et al. Classification of unmedicated bipolar disorder using whole-brain functional activity and connectivity: a radiomics analysis. Cereb Cortex 2020;30(3):1117-28. https://doi.org/10.1093/cercor/bhz152
16. Achalia R, Sinha A, Jacob A, Achalia G, Kaginalkar V, Venkatasubramanian G, et al. A proof of concept machine learning analysis using multimodal neuroimaging and neurocognitive measures as predictive biomarker in bipolar disorder. Asian J Psychiatr 2020;50:101984. https://doi.org/10.1016/j.ajp.2020.101984
17. Santos-Mayo L, San-Jose-Revuelta LM, Arribas JI. A computer-aided diagnosis system with EEG based on the P3b wave during an auditory odd-ball task in schizophrenia. IEEE Trans Biomed Eng 2017;64(2):395-407. https://doi.org/10.1109/tbme.2016.2558824
18. Lin GM, Nagamine M, Yang SN, Tai YM, Lin C, Sato H. Machine learning based suicide ideation prediction for military personnel. IEEE J Biomed Health Inform 2020;24(7):1907-16. https://doi.org/10.1109/jbhi.2020.2988393
19. Choi SB, Lee W, Yoon JH, Won JU, Kim DW. Ten-year prediction of suicide death using Cox regression and machine learning in a nationwide retrospective cohort study in South Korea. J Affect Disord 2018;231:8-14. https://doi.org/10.1016/j.jad.2018.01.019
20. Zheng L, Wang O, Hao S, Ye C, Liu M, Xia M, et al. Development of an early-warning system for high-risk patients for suicide attempt using deep learning and electronic health records. Transl Psychiatry 2020;10(1):72. https://doi.org/10.1038/s41398-020-0684-2
21. Bin-Hezam R, Ward TE. A machine learning approach towards detecting dementia based on its modifiable risk factors. Int J Adv Comput Sci Appl 2019;10(8):1-9.
22. Chen Q, Zhang-James Y, Barnett EJ, Lichtenstein P, Jokinen J, D'Onofrio BM, et al. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: a machine learning study using Swedish national registry data. PLoS Med 2020;17(11):e1003416. https://doi.org/10.1371/journal.pmed.1003416
23. Miche M, Studerus E, Meyer AH, Gloster AT, Beesdo-Baum K, Wittchen HU, et al. Prospective prediction of suicide attempts in community adolescents and young adults, using regression methods and machine learning. J Affect Disord 2020;265:570-8. https://doi.org/10.1016/j.jad.2019.11.093
24. Shen Y, Zhang W, Chan BS, Zhang Y, Meng F, Kennon EA, et al. Detecting risk of suicide attempts among Chinese medical college students using a machine learning algorithm. J Affect Disord 2020;273:18-23. https://doi.org/10.1016/j.jad.2020.04.057
25. Alimardani F, Cho JH, Boostani R, Hwang HJ. Classification of bipolar disorder and schizophrenia using steady-state visual evoked potential based features. IEEE Access 2018;6:40379-88. https://doi.org/10.1109/ACCESS.2018.2854555
26. Çolakoglu N, Akkaya B. Comparison of multi-class classification algorithms on early diagnosis of heart diseases. Proceedings of tge y-BIS Conference 2019: Recent Advances in Data Science and Business Analytics; 2019 Sep 25-28; Istanbul, Turkey. p. 162-71.
27. Department of Mental Health, Ministry of Public Health. Annual report 2020 [Internet]. Nonthaburi, Thailand: Ministry of Public Health; 2020 [cited at 2022 Oct 26]. Available from: https://www.dmh.go.th/report/dmh/rpt_year/dl.asp?id=461.
28. Onishi K. Risk factors and social background associated with suicide in Japan: a review. Jpn Hosp 2015;(34):35-50.
29. Gulabutr V. Suicide risk factors of Royal Thai Police Officers. Int J Crime Law Soc Issues 2017;4(2):65-80. https://doi.org/10.2139/ssrn.3261769