Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 219 (2023) 1279–1288

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN – International Conference on Project MANagement / HCist – International Conference on Health and Social Care Information Systems and Technologies 2022

An Apriori Algorithm-Based Association Rule Analysis to detect Human Suicidal Behaviour

Md. Mehedi Hassan[a], *Asif Karim[b]*, Swarnali Mollick[c], Sami Azam[d], Eva Ignatious[e], A S M Farhan Al Haque[f]

[a]*Computer Science and Engineering North Western University, Khulna, Bangladesh, mehedihassan@ieee.org*
[b]*College of Engineering, IT and Environment, Charles Darwin University, NT, Australia, asif.karim@cdu.edu.au*
[c]*Computer Science and Engineering Northern University of Business and Technology, Khulna, Bangladesh, swarnalimollick.07@gmail.com*
[d]*College of Engineering, IT and Environment, Charles Darwin University, NT, Australia, sami.azam@cdu.edu.au*
[e]*College of Engineering, IT and Environment, Charles Darwin University, NT, Australia, eva.ignatious@cdu.edu.au*
[f]*Department of Electronic Systems, Aalborg University Copenhagen, Denmark, Ahaque21@student.aau.dk*

* Corresponding author. *E-mail address:* asif.karim@cdu.edu.au

Abstract

Suicide is a major cause of death. It is also a complex public health issue and often preventable with timely intervention. Overall, the rate of suicide is increasing for various reasons. In our study, we use an association rule analysis to find the most important rules to predict suicidal behavior from an available data set. One of the most powerful machine learning algorithms available for identifying associations within databases is the Apriori algorithm. We used this algorithm to analyze association rules of suicidal behavior using a dataset of 1250 instances and 27 impactful features. These include daily activities, family background, and answers to mental questionnaires and have been analyzed to find combinations that are associated with suicidal behavior. The study has resulted in some key rules for human suicidal behavior. The Apriori method has been used to identify the eight most significant rules with the support of 0.25 and the confidence of 0.90.

© 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2022

Keywords: Suicide; Suicidal Behaviour; Behaviour Analysis; Association Rule Mining; Apriori.

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: asif.karim@cdu.edu.au

1877-0509 © 2023 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the CENTERIS – International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2022
10.1016/j.procs.2023.01.412

1. Introduction

Globally, suicide is a major public health concern. It is considered the second leading cause of death in young people (15–29 years old) and is a complex clinical challenge. It often relates to depression [1]. Socially created discrimination in power, responsibilities, roles, and status interacts with biological differences, which can lead to suicidal behavior. While considering the mental disorders including bipolar disorder, personality disorders, depression, schizophrenia, autism spectrum disorders, anxiety disorders, and physical disorders along with chronic fatigue syndrome, substance abuse such as alcoholism, and the use of benzodiazepines, stigma can be a risk factor in suicidal tendencies [2]. A major cause is depression. 95% of people who attempt suicide, including both fatal and non-fatal attempts, can be diagnosed with mental disorders, and in 80% of cases, the cause of suicide is depression [3]. Worldwide, the suicide rate is approximated at about 10.7 per hundred thousand. 10–20 million people try to commit suicide each year, and 50–120 million are deeply affected by the suicide or attempt to commit suicide of a close relative or close friend. Globally, the ratio of suicide deaths between males and females is 1.7: 1 [4]. However, there are regional differences. In Europe, the male-female suicide ratio is the highest, at 4.0, while it is the lowest in the Eastern Mediterranean region, at 1.1 [5]. On a global scale, females attempt two to four times more suicide attempts than males, most commonly between the ages of 15–19 years. Low and middle-income countries, contribute to around 79% of the world's total number of suicides [6]. Asia accounts for the majority of suicides occurring worldwide. Specifically, about 60% of the world's suicides happen in Asia. Each year, at least 60 million people in Asia are influenced by suicide attempts or attempt to commit suicide themselves. In our study, we attempt to identify major factors contributing to suicidal behavior. Data mining [7] methods are gaining broad application in different fields of medicine. This includes data cleaning, data integration, data selection, data transformation, and knowledge representation [8]. Association rule analysis is used as a practical approach for exploring warning symptoms and identifying high-risk people. For accurate prediction, the data needs to be collected periodically. The objective of our work is to identify the most critical rules for identifying suicidal people, to explore the main reasons for suicidal behavior, and to find possible correlations among these reasons for suicide.

2. Related Work

Amin et al. (2017) [9] developed a model using Artificial Neural Networks (ANN) [28] and Support Vector Machines (SVM) [29] to identify suicide causes in India. The accuracy for the ANN model was 77.5% and for the SVM model was 81.5%, which makes SVM a better option compared to ANN. The model used in their research was able to predict the number of suicides and identify the causes for all age groups independently, together with a male and female cause classification. By using data mining techniques, Joseph et al. (2018) [10] created a prediction strategy for individual people who are at a high risk of suicide. They studied the different suicide predictors, such as anxiety, depression, hopelessness, stress, etc. For the social network Twitter, they used several data mining algorithms such as Regression, Logistic Regression. Random Forest, Decision Table, and SMO to classify and identify tweets indicating a risk of suicide. The regression algorithm performed best and obtained the highest accuracy. The WEKA software was used to analyze the data. Tadesse et al. (2019) [11] demonstrated the potential of an LSTM-CNN model. They implemented data preprocessing by employing the Natural Language Toolkit (NLTK) and used machine learning with the NLP technique for feature extraction. A word embedding technique was used for deep learning. The LSTM-CNN model with word2vec features was used to classify the data based on text posts, links, or voting mechanisms for the posts on Reddit social media, by separating suicidal and non-suicidal posts. The accuracy rate of LSTM-CNN was 93.8% and the F1 score was 93.4% with optimized parameters. They evaluated the approach with four baseline methods, RF, SVM, NB, and XGBoost, and then proposed a model. The best outcome was obtained with LSTM-CNN. Ryu et al. (2018) [12] used public health data and applied a machine-learning algorithm to propose a strategy that can be used to identify individuals with suicidal ideation in the general population. While predicting suicide ideators in the test data set, the machine learning model demonstrated an accuracy of 78.3%. Teja et al. (2018) [13] used several machine learning (ML) algorithms to classify the data by taking factors such as age, cause, state, and the year of the incident into consideration. They used Decision Tree, Random Forest, Naıve Bayes, Logistic Regression, and AdaBoost algorithms for the classification. For the Decision Tree, the accuracy rate was only 49.07%, but an accuracy of 93.18%

was found for the Random Forest method. The accuracy for Logistic Regression, Random Forest, and AdaBoost were 78.54%, 81.67%, and 85.67%, respectively.

3. Methodology

Fig. 1 depicts the overall framework of our study and the various steps involved. These steps include Dataset Description, Data Preprocessing, Applying Association Rules, and Rule Extraction and Compare the Results..

Fig. 1. Workflow of the study

3.1. Dataset Description

Proper dataset selection and analysis make the study more relevant. We have used a dataset from Kaggle [14], consisting of 27 features and 1256 total instances, as shown in Table 1.

Table 1. Features List

Features	Data Distributions	
	Sub-category	Frequency
Age	29	85
	32	82
	26	75
	27	71
	33	70
	28	68
	Others	808
Gender	Male	937
	Female	221
	Others	101
Country	United States	751
	United Kingdom	185
	Canada	72
	Germany	45
	Ireland	27
	Netherlands	27
	Other	152
Self Employed	Yes	146
	No	1095
Family History	Yes	492
	No	767
Treatment	Yes	637
	No	622

Work Interfere	Never	213
	Often	144
	rarely	173
	sometimes	465
No Employees	100-500	176
	25-Jun	290
	26-100	289
	5-Jan	162
	500-1000	60
	More than 1000	282
Remote Work	Yes	376
	No	883
Tech Company	Yes	1031
	No	228
Benefits	Yes	477
	No	374
	Don't know	408
Care Options	Yes	444
	No	501
	Not Sure	314
Wellness Program	Yes	229
	No	842
	Don't know	188
Seek Help	Yes	250
	No	646
	Don't know	363
Anonymity	Yes	375
	No	65
	Don't know	819
Leave	Very easy	206
	Very difficult	98
	Somewhat easy	266
	Somewhat difficult	126
	Don't know	563
Mental Health Consequence	Yes	292
	No	490
	May be	477
Phys Health Consequence	Yes	61
	No	925
	May be	273
Coworkers	Yes	225
	No	260
	Some of them	774
Supervisor	Yes	516
	No	393
	Some of them	350
Mental Health Interview	Yes	44
	No	1008
	May be	207
Phys Health Interview	Yes	202
	No	500
	May be	557
Mental vs Physical	Yes	343
	No	340
	Don't know	576
Obs Consequence	Yes	184
	No	1075

3.2. Data Preprocessing

In data analysis, preprocessing is an essential step to modify the data for further processing. As an initial step, we have cleaned the dataset and handled the missing values. However, the median() function was used to compute most values.

3.3. Applying Association

A popular technique, association rule mining, is used for detecting the correlation among items in a data set. Association rule mining can be used to determine the patterns and frequent items in the data set. This process consists of two steps: 1) frequent items determination, 2) rules creation. The Apriori algorithm is used in this study as an unsupervised learning technique. The Apriori algorithm is used for the extraction of information from the dataset while at the same time association rule mining is performed, related to the requirements of the research. The Apriori algorithm is capable of mining item sets and creating association rules from the data set [15]. "Support" and "confidence" are two parameters used in this algorithm. Support indicates the frequently occurring items or the amalgamation of these items. Confidence is a conditional probability, which indicates how often items occur together. This is implemented by detecting frequent distinct items in the dataset and by extending the items to larger and larger item sets until enough items of the dataset are included in the item sets. Suppose, X and Y are two item sets. If the association is to be determined between them, then the minimum support and minimum confidence should be indicated. For our study, we are setting the minimum support to 25% and the minimum confidence to 90%. The frequent itemsets which are determined by Apriori can then be used to create association rules which highlight common trends in the data set. Association Rule Mining is a concept, mainly used for Basket Analysis for marketing purposes. The algorithm consists of two parts. In a data set, first get all the periodic groups with k number of items. This periodic k-itemset helps to use the self-join rule in order to find periodic groups with k+1 items. The algorithm works in an iterative way. Most of the association rules are created in the IF-THEN format. An association rule consists of an antecedent (if) and a consequent (then). It is used for finding relationships among variables in a large data set [16]. To measure association, three common metrics are defined below:

1) Support: The support of the rule is the prior probability of X and Y [17], The equation of support is :

$$s_{supp}(X \rightarrow Y) = \frac{(|X \cup Y|)}{n} \tag{1}$$

Here, "s_{supp}" is the support. "n" is the total number of transactions.

2) Confidence: The antecedent is the conditional probability of occurrence of the consequent. The confidence c of the rule is the conditional probability of Y given X [18]. Here, $\boldsymbol{C}$ = Confidence.

$$c(X \rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} \tag{2}$$

3) *Lift:* How important a rule is, measured by lift value. Basically, lift range can be defined by the filters of the rule. Life is found by the division of the actual confidence value and the expected confidence value of the rule [19]. Here, L = Lift.

$$L(X \rightarrow Y) = \frac{(T\ contains\ both\ X\ and\ Y)(T\ contains\ X)}{F\ of\ T\ contains\ Y} \tag{3}$$

Here, T= Transactions, F= Function.

3.4. Tools and techniques

The open-source language "R" is used for programming and analysis. "R" is a popular programming language for development. R is mainly an environment for statistical and graphical purposes. "R" is suitable for time series, machine learning [20], statistical inference, linear regression, and many other tasks [21, 30]. Some important functions are given below:

1) subset() [22]**:** When a particular condition is met, we get a data frame, vector, etc. from this function.
2) mean() [23]**:** This function is used to use the trim- med-mean.
3) chisq.test() [24]**:** As we need to test the chi-Square value, for this purpose we have used this function.
4) Apriori() [25]**:** Association mining rule is provid- ed by this function.
5) plot() [26]**:** A generic object to plot the objects of R.
6) inspect() [27]**:** All the summaries of the plot, option, and statistics are represented by this function.

4. Outcomes

We have tried to find impactful rules from the dataset and have extracted eight significant rules. We have used support of 25% and confidence of 90% for extracting our rules. These eight rules have been depicted in Fig. 2.

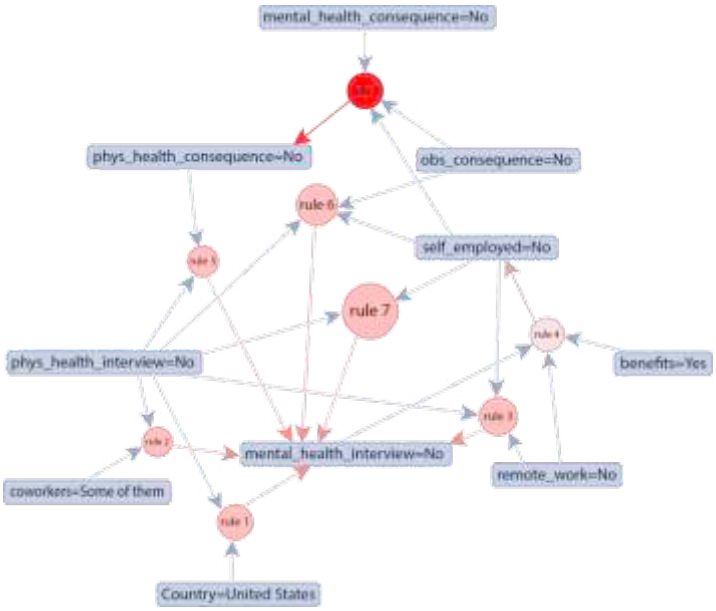

Fig. 2. Group of 8 significant rules

We have obtained the first rule where the country was the United States and the person was not willing to share his physical health issues in an interview. The person is not willing to share his mental health condition with a job interviewer with the support of 27% and the confidence of 99%. In Table 2, we have provided the details of all eight significant rules.

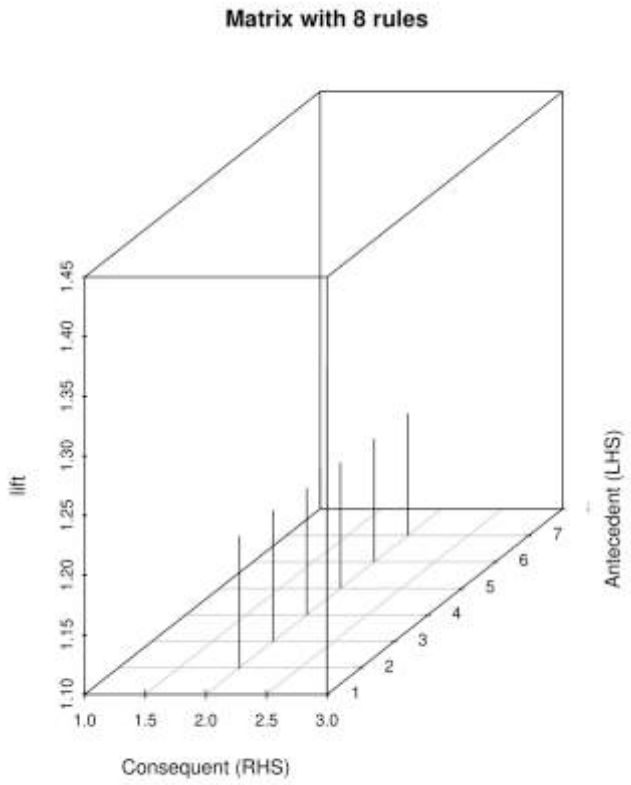

Fig. 3. Matrix with 8 rules

Table 2. Association Rules

Rules	LHS	RHS	Support	Confidence	Coverage	Lift	Count
[1]	{Country=United States,phys_health_interview=No}	{mental_health_interview=No}	0.2702	0.9925	0.2723	1.2106	264
[2]	{coworkers=Some of them,phys_health_interview=No}	{mental_health_interview=No}	0.2508	0.9919	0.2528	1.2098	245
[3]	{self_employed=No,remote_work=No,phys_health_interview=No}	{mental_health_interview=No}	0.2702	0.9888	0.2733	1.2060	264
[4]	{remote_work=No,benefits=Yes,mental_health_interview=No}	{self_employed=No}	0.2681679	0.9887	0.2712	1.1337	262
[5]	{phys_health_consequence=No,phys_health_interview=No}	{mental_health_interview=No}	0.2528	0.9880	0.2559	1.2051	247
[6]	{self_employed=No,phys_health_interview=No,obs_consequence=No}	{mental_health_interview=No}	0.2958	0.9863	0.2999	1.2031	289
[7]	{self_employed=No,phys_health_interview=No}	{mental_health_interview=No}	0.3469	0.9855	0.3521	1.2019	339
[8]	{self_employed=No,mental_health_consequence=No,obs_consequence=No}	{phys_health_consequence=No}	0.2712	0.9851	0.2753	1.3749	265

In the plot of Fig. 3, the feature ‘lift’ is displayed. The ranking of the rules is based on the lift. Fig. 4 demonstrates the confidence and Fig. 5 demonstrates the parallel coordinate plot of these eight rules.

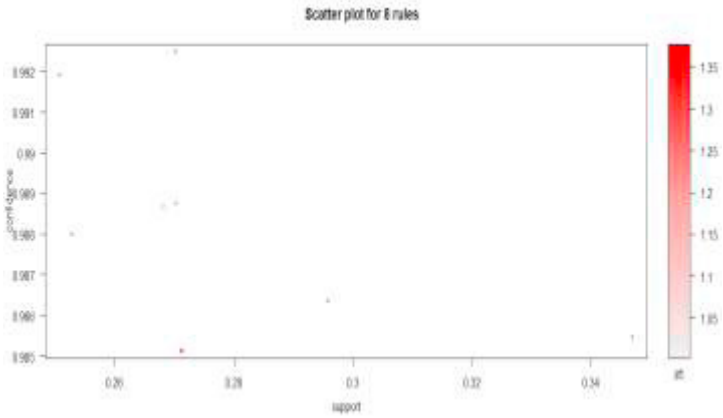

Fig. 4. Scatter plot of 8 rules.

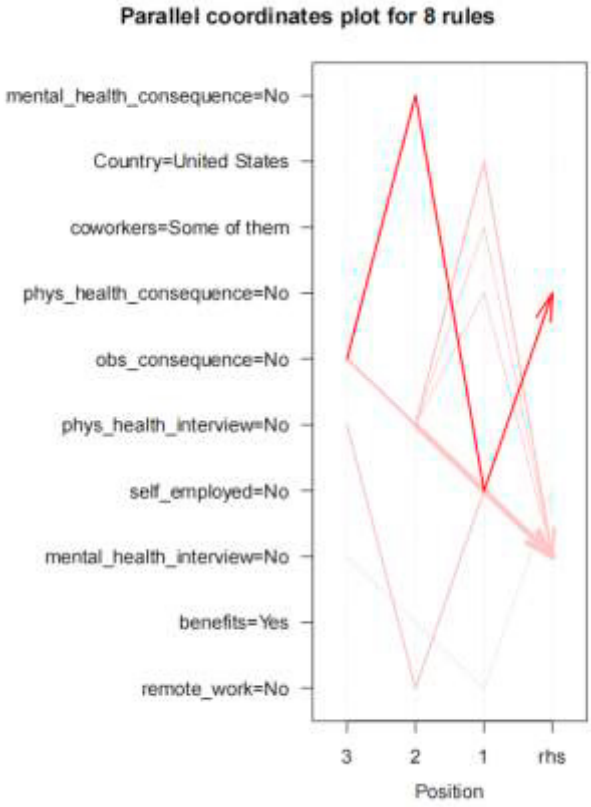

Fig. 5. Parallel coordinates plot for 8 rules.

5. Conclusion

We have analyzed factors related to suicide attempts. From the available dataset, we identified some key factors and interconnections between them. We have used the Apriori algorithm for analyzing the relationships by training the algorithm with the dataset. Our research found eight association rules with 25% is support and the confidence of 90%. In future work, we intend to develop an Android application for calculating the suicide risk level, especially for people who are depressed.

References

[1] Bilsen, J. (2018) "Suicide and Youth: Risk Factors." *Frontiers in Psychiatry* 9.

[2] Murri, M., Gancitano, M., Antenora, F., Mojtahedzadeh, M., Salman, J. (2021) "Depression and Substance Use Disorders in Physicians." *Depression, Burnout and Suicide in Physicians* 37-53.

[3] Vogelzang, B., Scutaru, C., Mache, S., Vitzthum, K., Quarcoo, D., Groneberg, D. (2011) "Depression and Suicide Publication Analysis, Using Density Equalizing Mapping and Output Benchmarking." *Indian Journal of Psychological Medicine* 33(1): 59-65.

[4] Albert, P. (2015) "Why is depression more prevalent in women?." *Journal of Psychiatry and Neuroscience* 40(4): 219-221.

[5] Värnik, P., (2012) "Suicide in the World." *International Journal of Environmental Research and Public Health* 9(3): 760-771.

[6] Beattie, T., Smilenova, B., Krishnaratne, S., Mazzuca, A. (2020) "Mental health problems among female sex workers in low- and middle-income countries: A systematic review and meta-analysis", *PLOS Medicine* 17(9): e1003297.

[7] M. Billah, M. Hassan, M. Raihan, S. Mollick, M. Hasan Shakil, J. Angon. (2021). "A Data Mining Approach to Identify Human Behavior on Different Activities of Student." *12Th International Conference On Computing Communication And Networking Technologies (ICCCNT)*.

[8] Boonkwang, K., Kasemvilas, S., Kaewhao, S., Youdkang, O. (2018) "A Comparison of Data Mining Techniques for Suicide Attempt Characteristics Mapping and Prediction." *2018 International Seminar on Application for Technology of Information and Communication.*

[9] Syed, S., Amin, I. (2017) "Prediction of Suicide Causes in India using Machine Learning." *Journal of Independent Studies and Research - Computing* 15(2).

[10] Joseph, A., Ramamurthy, B. (2018). "Suicidal behavior prediction using data mining techniques", *Int J Mech Eng Technol* 9(4): 293-301.

[11] Tadesse, M., Lin, H., Xu, B., Yang, L. (2019) "Detection of Suicide Ideation in Social Media Forums Using Deep Learning." *Algorithms* 13(7): 7.

[12] Ryu, S., Lee, H., Lee, D., Park, K. (2018) "Use of a Machine Learning Algorithm to Predict Individuals with Suicide Ideation in the General Population." *Psychiatry Investigation* 15(11): 1030-1036.

[13] Teja K., Pravalika S., Varshitha G., Basha S. (2018) "Classification of Suicidal Deaths Caused in India through Various Supervised Machine Learning Techniques." *2018 International Journal of Engineering Research in Computer Science and Engineering* 5: 237-242

[14] Patel P. (2018) "Perceived Workplace Fa ctors and their Influence on Self-Reported Mental Health Service Seeking Among Technology Workers".

[15] Jena, L., Kamila, N. (2014) "A Model for Prediction of Human Depression Using Apriori Algorithm." *2014 International Conference on Information Technology*.

[16] Hassan, M., Peya, Z., Zaman, S., Angon, J., Keya, A., Dulla, A. (2020) "A Machine Learning Approach to Identify the Correlation and Association among the Students' Drug Addict Behavior." *2020 11th International Conference on Computing, Communication and Networking Technologies* (ICCCNT) 1-5.

[17] Chen, G., Liu, H., Yu, L., Wei, Q., Zhang, X. (2006) "A new approach to classification based on association rule mining." *Decision Support Systems* 42(2): 674-689.

[18] Zhong, Y., Liao, Y., 2012. Research of Mining Effective and Weighted Association Rules Based on Dual Confidence. 2012 Fourth International Conference on Computational and Information Sciences.

[19] McNicholas, P., Murphy, T., O'Regan, M. (2008) "Standardising the lift of an association rule." *Computational Statistics & Data Analysis* 52(10): 4712-4721.

[20] M.S. Junayed, A.A. Jeny, S.T. Atik, N. Neehal, A. Karim, S. Azam, et al. (2019), "AcneNet - a deep CNN based classification approach for Acne Classes", 2019 12th *International Conference on Information and Communication Technology and System (ICTS)*. doi:10.1109/icts.2019.8850935.

[21] Raihan, M., Islam, M., Ghosh, P., Hassan, M., Angon, J., Kabiraj, S. (2020) "Human Behavior Analysis using Association Rule Mining Techniques." *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*.

[22] Hothorn, T., Everitt, B. (2006) A Handbook of Statistical Analyses Using R.

[23] Gelman, A. (2008) "Scaling regression inputs by dividing by two standard deviations." *Statistics in Medicine* 27(15): 2865-2873.

[24] Li, Y., Baron, J. (2011) "Statistics for Comparing Means and Proportions." *Behavioral Research Data Analysis with R* 39-54.

[25] Mehta, A., Bura, D. (2020) "Mining of Association Rules in R Using Apriori Algorithm." *Lecture Notes in Electrical Engineering* 181-188.

[26] Paradis, E., Claude, J., Strimmer, K. (2004) "APE: Analyses of Phylogenetics and Evolution in R language." *Bioinformatics* 20(2): 289-290.

[27] Anders, S., McCarthy, D., Chen, Y., Okoniewski, M., Smyth, G., Huber, W., Robinson, M. (2013) "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor." *Nature Protocols* 8(9): 1765-1786.

[28] Ghosh, P., Azam, S., Karim, A., Jonkman, M., Hasan, M. (2021) "Use of Efficient Machine Learning Techniques in the Identification of Patients with Heart Diseases." *2021 the 5th International Conference on Information System and Data Mining.*

[29] Mahmud, K., Azam, S., Karim, A., Zobaed, S., Shanmugam, B., Mathur, D. "Machine learning based PV Power Generation Forecasting in Alice Springs", *IEEE Access*. 9 (2021) 46117–46128. doi:10.1109/access.2021.3066494.

[30] Ahmad, M., Nourzadeh, H., Siebers, J., A regression‐based approach to compute the pixels sensitivity map of Linear Accelerator Portal Imaging Devices, *Medical Physics.* 48 (2021) 4598‐4609. doi:10.1002/mp.14862.