A Flexible Semi-Synthetic Data Generator for Risky Drinking Behavior

Chi Hao Liow
Korea Advanced Institute of Science and Technology (KAIST)
Youngwoo Choi
Korea Advanced Institute of Science and Technology (KAIST)
Jiwon Yeom
Korea Advanced Institute of Science and Technology (KAIST)
Young Yim Doh
Korea Advanced Institute of Science and Technology
Seungbum Hong (seungbum@kaist.ac.kr)
Korea Advanced Institute of Science and Technology (KAIST)

Article

Keywords: data scarcity, synthetic data, data augmentation, binge-drinking behavior, survey

Posted Date: March 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1425863/v1

Abstract

Machine intelligence has garnered immense attention owing to its ability to discover hidden patterns in abstract and high-dimensional datasets. However, its success is often limited by the fundamental bottleneck of data scarcity. In this work, we offer a universal data augmentation solution to resolve this impasse. We first discovered the hidden knowledge within the existing scarce dataset using the machine learning (ML) technique and then synthetically augmented the dataset according to its feature importance. In principle, scarce and augmented datasets should share a common statistical property. Using this property, we specifically study the scarce dataset representing the binge-drinking behavior of university students and show that our method is effective in augmenting a limited dataset with high fidelity. The current work challenges the status quo in data scarcity with rule-less-based ML, which removes the ostensible barrier that prevents the application of data-driven techniques to the data scarce clinical research.

Introduction

Binge drinking (BD) is often associated with an increased risk of traumatic injuries, ischemic heart disease, and adverse psychological consequences[1–3]. It is estimated that more than 90% of Korean college students are engaged in drinking, with 20–37% of them being binge drinkers[4–6]. This alarming rate of BD behavior among college students has called upon an urgent intervention before students develop a strong alcohol dependency.

Previous research works have used the theory of planned behavior (TPB) as the basic survey framework to investigate the correlation of students' motivation to drinking behavior[6–8]. For example, Norman *et al*. showed that social environments, such as peer pressure and positive environment, strongly influence students' drinking behavior[8]. Because of the flexibility of TPB, Chen *et al*. added a negative "stress" variance into TPB and concluded a positive correlation between stress and BD[6]. In addition, Lannoy *et al*. pointed out that positive emotions, such as happiness, can also be an important indicator of BD[9–11]. However, because human behaviors are dynamic, they may change according to the surrounding environment context, peers, or internal emotional state[12]. Therefore, a rapid and reliable prognostic tool is required to analyze the cognitive factors that motivate students' BD behavior, thus helping to design a usual intervention strategy.

With the ability to process a large number of complex datasets at a fast pace, machine learning (ML) has recently become a reliable and critical instrument for data analysis and has permeated a wide range of domains[13–16], including natural language processing[17], digital safety[18], games[19], and finance[20]. Likewise, excitement and attention have also been drawn into the field of behavioral sciences, aiming to identify causal underpinnings that govern human behaviors, which can be performed through facial expression recognition[21, 22], speech classification[23, 24], and electroencephalogram screening[25, 26]. In a

recent study, by training on more than 45,000 data, Kim *et al*. used ML to determine the leading factors to participants' drinking problem, which later helped facilitate a proper treatment plan[27].

All the examples cited above use ML for behavior analysis driven by an unlimited stream of data. Ironically, many systems of interest are bound to the fundamental issue of data scarcity. Therefore, the application of ML in these systems remains a major challenge. A few strategies have recently been proposed to circumvent the data limitation issue. For example, by adapting the transfer-learning method, one can first pretrain the proxy-related properties of an adequate dataset and subsequently transfer the "knowledge" to another domain where data are scarce[28, 29]. Another approach is to create an inexhaustible amount of simulated data governed by a physical law[30–32]. However, rule-less-based research, such as survey methodology, does not generally fit into any of the above criteria. Arguably, a robust method guided by ML synthetic data augmentation could be a parsimonious solution to this issue.

Inspired by the generative adversarial network method that generates new data with the same statistical characteristics as their training dataset[33], we attempt to develop a universal semi-synthetic data generator (SSG) that tackles the principle bottleneck of data scarcity for college students' BD behavior at the Korea Advanced Institute of Science and Technology (KAIST). We first inputted the scarce dataset into ML as training data. Subsequently, we augmented the dataset based on the statistical property evaluated by feature importance (FI). The SSG model consists of three parts: (1) data collection related to students' BD behavior based on the extended TPB, (2) ML model selection based on the evaluation of the robustness over systematic error and computation cost, (3) integrating the filtering strategy (FS) based on FI to ensure the quality of the augmented dataset. We demonstrate the SSG model's ability to capture the statistical relationship (~ 12% FI variance) with a reasonable validation accuracy (> 0.8). Further analysis using the decision tree (DT) algorithm indicates that the factor of "happiness" is more prominent in predicting the students' BD behavior than normally perceived stress-driven alcohol consumption. In accordance with the DT interpretation of the dataset, we deployed an interactive mobile-app-based chatbot to complete the design-to-device pipeline. This work establishes a viable universal model for augmenting scarce datasets and demonstrates unprecedented potential in data-driven applications.

Methods

Survey design and data collection. The survey questions were designed according to the framework of the Theory of Planned Behavior (TPB) developed by Ajzen to predict health/ risky behavior from a cognitive perspective[34–36]. In the extended version of TPB, we included happiness and stress, which emphasize the positive and negative role of emotions in adaptive and maladaptive behaviors. The happiness and stress were evaluated using standard Perceived Stress Scale Assessment and Subjective Happiness Scale[37, 38]. In this study, the respondent's *Attitude* towards binge drinking was measured by 8 items with 5-point Likert Scale (e.g. For me to engage binge-drinking next week will reduce my stress, the scales range from 1 (strongly disagree) to 5 (strongly agree) ($\alpha = 0.95$). *Subjective Norm* was measured by 6 items (e.g. My friends (or classmates) think I should have binge-drinking next week) ($\alpha = 0.84$).

Perceived Behavioral Control was measured with 3 items (e.g., I have total control of how much alcohol I drink every time) (α = 0.48). *Stress* was measured by 10 items (e.g. In the last month, how often have you been upset because of something that happened unexpectedly?), the response scales ranged 1 = never to 5 = very often. High values indicate high level of stress) (α = 0.47). *Happiness* was measured by 4 items (e.g. In the month, I consider myself, 1 = “not a very happy person” to 5 = “a very happy person”. High values indicate high level of happiness) (α = 0.47).

Two waves survey at 1-week interval were conducted to the students of Korea Advanced Institute of Science and Technology (KAIST). The survey was performed in accordance with the Bioethics and Safety Act, and the Declaration of Helsinki, and was approved by the Institutional Review Board of KAIST (IRB: KH2019-171). Informed consent was obtained from all subjects before they proceeded to complete the survey.

This study defined the binge-drinking according to the National Institute of Alcohol Abuse and Alcoholism[39], which is a pattern of episodic alcohol consumption (more than four (five) drinks for female (male)) in about two hours. Before the data augmentation, we categorized the collected data into two classes: binge-drinking and non-binge-drinking. The data shows an imbalanced distribution between the two classes, with the ratio of majority to minority being 7:32 (binge: non-binge-drinking). A Synthetic Minority Over-Sampling Technique (SMOTE) is typically used to balance the input data to avoid the model misclassification[40–42]. In SMOTE, a k-nearest neighbors (KNN) algorithm was used to calculate the Euclidean distance in generating synthetic data for over-sampling the minority sample. The resulting dataset with balanced classes is illustrated in Fig S2. Even after the SMOTE balancing, the amount of dataset is still not sufficient to represent the whole picture; therefore, data augmentation is needed. Henceforth, the balanced data will be used for all the subsequent data augmentation.

Concept of SSG. Figure 1 illustrates the concept of SSG model in this study. Metaphorically, a scarce dataset can be an analogy to puzzle pieces sparsely spread across problem-specific parameter spaces. By training the ML model with the given dataset, ML model is able to “learn” the hidden knowledge. To connect the missing pieces, the trained ML model predicts new temporary data (D_{temp}) by feeding randomly generated variables from predefined parameter spaces. A FS is applied to each D_{temp} to justify if the new prediction fit into the FI criteria. If D_{temp} fail to meet the criteria, it will be discarded. The cycle will continue until the desired amount of data is created.

Ideally, an effective SSG should offer the highest prediction accuracy and lowest computational cost. An essential key to achieving this objective is to employ a suitable ML model for the SSG. Here, we compared three different models: support vector machine (SVM), neural networks (NNs), and XGBoost (XGB). The accuracy is determined by the coefficient of determination (R^2) according to Eq. (1)[43]:

$$R^2 = 1 - \frac{\Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \bar{y})^2}, \qquad (1)$$

where y_i, $\hat{y}_i$, and $\bar{y}$ represent the i^{th} true value, i^{th} predicted value, and mean value of y, respectively.

Finally, we visualized the augmented data with decision tree diagram based on the optimized DT model. To demonstrate the potential application of this research, we deployed the trained DT model into an interactive mobile-app-based based on Telegram with the name of @DrinkingBehavior_bot.

Results

Survey analysis and SSG models evaluation. Here, we accounted for "Stress" and "Happiness" as the positive and negative emotional features in predicting BD behavior, as schematically illustrated in Fig. S1[6, 36]. The survey was separated into two waves with one-week intervals. Total 39 respondents had completed both surveys, among which 61.8% were male, and 38.2% were female students in the age range between 19 and 36.

The Cronbach's α in Table S1 represents the internal consistency between independent variables. We noticed that some variables (Stress, Happiness, and Perceived Control) yielded low Cronbach's α (0.47–0.95). Nevertheless, the internal consistency reliability of the present study is considered acceptable, especially given the small number of items in the survey[44, 45]. One can increase the Cronbach's α by increasing the number of items [46].

Figure 2 shows the training and validation accuracies derived from the optimized ML models. Based on the learning curves, as the number of data increased, the range of validation accuracy also increased from 0.65 and reached > 0.85 after the 30th iteration. This result signifies that a large amount of data is crucial in reducing misclassification and effectively generalizing complex and nonlinear data.

SVM outperformed XGB and NN models, as evidenced by the training and validation accuracy and computational time. As SVM relies only on the single hyperparameter C in fitting nonlinear data, it has less computational burden during the grid search process [47]. Contrary, the intensive hyperparameter optimization in sophisticated NN and XGB models require much more time to complete the task. Therefore, we employed the SVM model for subsequent SSG operations.

Data generation. Naturally, a trained ML model can provide a large number of predicted D_{temp}. Although it may increase the size of the scarce dataset, ensuring the quality of statistical relationships in our prediction can be challenging, particularly in describing problem-specific features. In principle, datasets that share similar information should exhibit similar statistical relationships and trends. Therefore, an FS design based on FI was rationally employed to ensure that the statistical relationship between features is preserved before and after the SSG operation. FI is a powerful technique for discovering nonlinear statistical relationships governing input features[48]. By predefining the FI variance between the standard and augmented data to be ≤ 20%, any unfit D_{temp} was removed to preserve each feature's relative contribution to the target. In this way, only D_{temp} that best describes the scarce dataset will remain.

Finally, the passed D_{temp} was updated into a scarce dataset and used for the next cycle of data augmentation. This cycle continued until the desired amount of dataset was achieved.

Further insight is provided in Fig. 3, which shows the FI analysis that quantifies the relative contribution of each feature to the BD behavior. The FI trends of the standard (scarce dataset) and SSG with the FS-generated dataset are similar, displaying the results in a descending order: happiness > subjective norm > stress > attitude > intention > perceived control. Conversely, the FI trend of SSG without the FS-generated dataset displayed the results in the following descending order: intention > happiness > subjective norm > attitude > stress > perceived control.

The FI variances between the SSG and standard dataset, which is calculated from $|FI_{std} - FI_{ssg}|/FI_{std}$ (where subscripts *std* and *ssg* represent the standard and SSG datasets, respectively), are ~ 36% (without FS) and ~ 12% (with FS). Because the SSG without FS is not bound to any constraint, the augmented dataset may result in a large variance with a random FI trend. Furthermore, the small FI variance with the FS implementation proves that it is an effective method to augment a representable dataset with high confidence rather than ad hoc regression. In addition, the SSG-augmented dataset's viability can be observed from their validation accuracy, as illustrated in Fig. S2, where all the selected D_{temp} yielded a validation accuracy greater than 0.8.

Discussion

With the SSG model, we augmented a scarce student's BD dataset, guided by their intrinsic statistical features. Subsequently, one can exploit the DT for its intuitive ability to extract hidden information from complex data in a way that is readily understood by a human. For comparison, Fig. 4a and Table S2 show the DT performance based on the receiver operating characteristic curves and F1 score trained on the augmented dataset. The augmented dataset enables the DT to generalize better than that of the scarce one, with area under the curve (AUC) values of 0.90 for the SSG (with FS) and 0.76 for the scarce dataset.

To explore the insights of the dataset, we visualized a tree diagram of the augmented dataset based on the optimized DT. In Fig. 4b, the topmost node with a Gini index of 0.5 belongs to the topmost node of "happiness," and it branches out into two child nodes that belong to "subjective norm" and "intention." The tree grows until it reaches a Gini index equal to zero (see https://github.com/lchlyw/SSG_augmentation for the fully-grown trees). Generally, features that are near the topmost node are better at classifying the BD behavior. As such, positive emotion (happiness) is the most important factor that predicts the students' BD behavior. The results show that the students are more likely to engage in BD when surrounded by an encouraging environment and a positive mood. In many ways, students might experience happy feelings, and one good example can be observed at the end of the academic term. With peers' encouragement/approval, it might be difficult to resist the temptation of excessive drinking, which is in agreement with the result of previous research related to enhancement motive and social-contextual factors[49, 50].

Contrary to the neuroscience perspective, negative emotions, such as stress, are often hypothesized as indicators of the BD behavior[6, 51]. According to the rodent alcohol dependency model, the brain releases the corticotropin-releasing factor signal during a BD session, which governs the desire for alcohol drinking and anxiety[51–53]. Such a negative reward-seeking behavior can further increase the negative emotional state and alcohol consumption.

However, the current study shows that stress is the third most important factor in BD behavior after happiness and subjective norms. We anticipated two reasons for our observation: (1) First, our survey period was conducted during the university holiday (June 2020), which was at the start of the summer semester. Therefore, students do not feel too much academic stress; hence, stress is less significant than happiness in determining their BD behavior. (2) Rather than a tool to cope with stress, students' alcohol drinking occasions are usually related to social activities, which are often motivated by positive enhancement.

We completed the design-to-device pipeline based on the DT diagram by demonstrating the knowledge transfer from complex and abstract data into an interactive mobile-app-based chatbot. Figure 4c illustrates an accessible Telegram chatbot (@DrinkingBehavior_bot) interface built with Dialogflow. According to the DT classification diagram, the chatbot design is based on a simple yes/no question–answer format related to the BD behavior. Because the current study aims to lay the foundation for building a viable problem-specific data augmentation model, a clinical trial is needed before it can be officially used as a professional diagnostic tool for mental health issues.

Limitations of the Current Work. Our study has several limitations. First, because our data were collected based on a single location in KAIST, it is important to collect data from other universities to validate our results. Second, although ML considers a black box that can uncover a hidden pattern within the dataset, it is always difficult to extrapolate into an unseen region. Therefore, our SSG system can only generate new data within preset boundaries with statistical characteristics similar to their training dataset. Third, to increase the Cronbach's α, we should modify the questions in perceived stress scale assessment, subjective happiness scale, and perceived control and to capture a wider variance of behavior, more feature vectors should be included in the data collection process.

Future efforts can be made by administering this system in an actual clinical trial where we can compare the generated data with that of an actual observation, to validate the accuracy of SSG-generated data. Based on the proposition that students' emotional state and social norms play an important role in motivating students into BD, an intervention design should focus on students' emotional states and social norms.

Conclusions

Overall, scarce problem-specific data augmentation has been achieved via the judicious application of the SSG model. This study is divided into three steps: First, we conducted an online survey related to the

BD behavior of university students. Second, we selected the SVM model to be integrated into the SSG model because of its high validation accuracy and low computational cost. Third, by integrating FS into the SSG model to enhance the quality of the augmented dataset, we obtained a low FI variance of ~ 12% and high validation accuracy (> 0.8). Using the DT algorithm, the augmented dataset exhibited a high validation AUC (0.90) compared to that (0.76) of the scarce dataset. In addition, this study suggests that "happiness" is the strongest predictive feature for students' BD behavior. For a proof of concept, we further explored translating abstract knowledge into an interactive design-to-device pipeline on a mobile-apps-based chatbot. The current work allows for realization of data-driven application and remove the "curse" of data scarcity. Replacing the target input data from binge-drinking with other scarce clinical data could enable faster and more accurate integration of data-driven technique into their fields.

Declarations

Acknowledgments

This work was supported by the KAIST-funded Global Singularity Research Program for 2020, 2021 and 2022, and the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5C2A03081332). We gratefully acknowledge Prof. Meeyoung Cha at KAIST for her insightful discussion on the machine learning part.

Authors Contributions: C.L. and S.H. designed the survey and machine learning framework. C.L., Y.C., and J.Y. conducted the BD behavior survey. C.L. conducted the machine learning tasks. All the authors participated in the discussion of the data analysis. C.L., Y.Y.D., and S. H. wrote the manuscript.

Competing Interest Statement: The authors declare no competing financial interest.

This article contains supporting information.

Data availability: All data and codes are available at https://github.com/lchlyw/SSG_augmentation.

References

1. Griswold, M.G., et al. Alcohol use and burden for 195 countries and territories, 1990–2013; 2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. **392**, 1015–1035 (2018).
2. Ruidavets, J.B., P. Ducimetiere, A. Evans, M. Montaye, B. Haas, A. Bingham, J. Yarnell, P. Amouyel, D. Arveiler, F. Kee, V. Bongard, and J. Ferrieres Patterns of alcohol consumption and ischaemic heart disease in culturally divergent countries: the prospective epidemiological study of myocardial infarction (PRIME). BMJ. **341**, c6077 (2010).
3. Schaffer, M., E.L. Jeglic, and B. Stanley The relationship between suicidal behavior, ideation, and binge drinking among college students. Arch. Suicide Res. **12**, 124–132 (2008).
4. Park, J.M., A. Sohn, and C. Choi Solitary and social drinking in South Korea: An Exploratory Study. Osong Public Health Res. Perspect. **11**, 365–372 (2020).

5. Chung, H.-K. and H.-Y. Lee Drinking behaviors by stress level in Korean university students. Nutr. Res. Pract. **6**, 146–154 (2012).
6. Chen, Y. and T.H. Feeley Predicting binge drinking in college students:Rational beliefs, stress, or loneliness? J. Drug Educ. **45**, 133–155 (2015).
7. Norman, P. The theory of planned behavior and binge drinking among undergraduate students: Assessing the impact of habit strength. Addict. Behav. **36**, 502–507 (2011).
8. Norman, P., P. Bennett, and H. Lewis Understanding binge drinking among young people: an application of the theory of planned behaviour. Health Educ. Res. **13**, 163–169 (1998).
9. Lannoy, S., V. Dormal, M. Brion, J. Billieux, and P. Maurage Preserved crossmodal integration of emotional signals in binge drinking. Front. Psychol. **8**, (2017).
10. Lannoy, S., L. Dricot, F. Benzerouk, C. Portefaix, S. Barriere, V. Quaglino, M. Naassila, A. Kaladjian, and F. Gierski Neural responses to the implicit processing of emotional facial expressions in binge drinking. Alcohol Alcohol. **56**, 166–174 (2021).
11. Lannoy, S., T. Duka, C. Carbia, J. Billieux, S. Fontesse, V. Dormal, F. Gierski, E. Lopez-Caneda, E.V. Sullivan, and P. Maurage Emotional processes in binge drinking: A systematic review and perspective. Clin. Psychol. Rev. **84**, (2021).
12. Schill, C., J.M. Anderies, T. Lindahl, C. Folke, S. Polasky, J.C. Cardenas, A.S. Crepin, M.A. Janssen, J. Norberg, and M. Schluter A more dynamic understanding of human behaviour for the Anthropocene. Nat. Sustain. **2**, 1075–1082 (2019).
13. Kohonen, T., E. Oja, O. Simula, A. Visa, and J. Kangas Engineering applications of the self-organizing map. Proc. IEEE. **84**, 1358–1384 (1996).
14. Kourou, K., T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, and D.I. Fotiadis Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015).
15. Libbrecht, M.W. and W.S. Noble Machine learning applications in genetics and genomics. Nat. Rev. Genet. **16**, 321–332 (2015).
16. Ramprasad, R., R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim Machine learning in materials informatics: recent applications and prospects. npj Comput. Mater. **3**, 54 (2017).
17. Tshitoyan, V., J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K.A. Persson, G. Ceder, and A. Jain Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. **571**, 95–98 (2019).
18. Koopman, P. and M. Wagner Autonomous vehicle safety: An interdisciplinary challenge. IEEE Intell. Transp. Syst. Mag. **9**, 90–96 (2017).
19. Silver, D., et al. Mastering the game of Go with deep neural networks and tree search. Nature. **529**, 484-+ (2016).
20. Meng, T.L. and M. Khushi Reinforcement Learning in Financial Markets. Data. **4**, 110 (2019).
21. Rouast, P.V., M.T.P. Adam, and R. Chiong Deep learning for human affect recognition: Insights and new developments. IEEE Trans. Affect. Comput. **12**, 524–543 (2021).

22. Jaison, A. and C. Deepa A review on facial emotion recognition and classification analysis with deep learning. Biosci. Biotechnol. Res. Commun. **14**, 154–161 (2021).
23. Fayek, H.M., M. Lech, and L. Cavedon Evaluating deep learning architectures for speech emotion recognition. Neural Netw. **92**, 60–68 (2017).
24. Pandey, S.K., H.S. Shekhawat, and S.R.M. Prasanna. *Deep learning techniques for speech emotion recognition : A review*. in *29th International Conference on Radioelektronika (RADIOELEKTRONIKA) / Microwave and Radio Electronics Week (MAREW, 2019)*. 2019. Pardubice, CZECH REPUBLIC.
25. Barros, C., C.A. Silva, and A.P. Pinheiro Advanced EEG-based learning approaches to predict schizophrenia: Promises and pitfalls. Artif. Intell. Med. **114**, (2021).
26. de Bardeci, M., C.T. Ip, and S. Olbrich Deep learning applied to electroencephalogram data in mental disorders: A systematic review. Biol. Psychol. **162**, (2021).
27. Kim, S.-Y., T. Park, K. Kim, J. Oh, Y. Park, and D.-J. Kim A Deep Learning Algorithm to Predict Hazardous Drinkers and the Severity of Alcohol-Related Problems Using K-NHANES. Front. Psychiatry. **12**, (2021).
28. Wu, S., Y. Kondo, M.-a. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, and R. Yoshida Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. npj Comput. Mater. **5**, 66 (2019).
29. Yamada, H., C. Liu, S. Wu, Y. Koyama, S.H. Ju, J. Shiomi, J. Morikawa, and R. Yoshida Predicting materials properties with little data using shotgun transfer learning. ACS Cent. Sci. **5**, 1717–1730 (2019).
30. Seko, A., A. Togo, H. Hayashi, K. Tsuda, L. Chaput, and I. Tanaka Prediction of low-thermal-conductivity compounds with first-principles anharmonic lattice-dynamics calculations and bayesian optimization. Phys. Rev. Lett. **115**, 205901 (2015).
31. Cubuk, E.D., A.D. Sendek, and E.J. Reed Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data. J. Chem. Phys. **150**, 214701 (2019).
32. Chandrasekaran, A., D. Kamal, R. Batra, C. Kim, L.H. Chen, and R. Ramprasad Solving the electronic structure problem with machine learning. npj Comput. Mater. **5**, 22 (2019).
33. Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio Generative adversarial networks. Commun. ACM. **63**, 139–144 (2020).
34. Ajzen, I. Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior. J. Appl. Soc. Psychol. **32**, 665–683 (2002).
35. Ajzen, I. The theory of planned behaviour: Reactions and reflections. Psychol. Health. **26**, 1113–1127 (2011).
36. Ajzen, I. The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**, 179–211 (1991).
37. Cohen, S., T. Kamarck, and R. Mermelstein A Global Measure of Perceived Stress. J. Health Soc. Behav. **24**, 385–396 (1983).

38. Lyubomirsky, S. and H.S. Lepper A Measure of Subjective Happiness: Preliminary Reliability and Construct Validation. Soc. Indic. Res. **46**, 137–155 (1999).
39. National Institute of Alcohol Abuse and Alcoholism [NIAAA], *NIAAA Council approves definition of binge drinking*, in *NIAAA Newsletter*. 2004.
40. Goh, Y.M., C.U. Ubeynarayana, K.L.X. Wong, and B.H.W. Guo Factors influencing unsafe behaviors: A supervised learning approach. Accid. Anal. Prev. **118**, 77–85 (2018).
41. Moharreri, E., M. Pardakhti, R. Srivastava, and S.L. Suib Energy-Geometry Dependency of Molecular Structures: A Multistep Machine Learning Approach. ACS Comb. Sci. **21**, 614–621 (2019).
42. Blagus, R. and L. Lusa SMOTE for high-dimensional class-imbalanced data. BMC Bioinform. **14**, 106 (2013).
43. McCartney, M., M. Haeringer, and W. Polifke Comparison of machine learning algorithms in the interpolation and extrapolation of flame describing functions. J. Eng. Gas Turb. Power. **142**, 061009 (2020).
44. Itani, L., H. Chatila, H. Dimassi, and F. El Sahn Development and validation of an Arabic questionnaire to assess psychosocial determinants of eating behavior among adolescents: a cross-sectional study. J. Health Popul. Nutr. **36**, 10 (2017).
45. Cortina, J.M. What is coefficient alpha? An examination of theory and applications. J. Appl. Soc. Psychol. **78**, 98–104 (1993).
46. Lee, E.-H. Review of the psychometric evidence of the perceived stress scale. Asian Nurs. Res. **6**, 121–127 (2012).
47. Joshi, R.P., J. Eickholt, L.L. Li, M. Fornari, V. Barone, and J.E. Peraltata Machine learning the voltage of electrode materials in metal-ion batteries. ACS Appl. Mater. Interfaces. **11**, 18494–18503 (2019).
48. Li, Z., S.W. Wang, W.S. Chin, L.E. Achenie, and H.L. Xin High-throughput screening of bimetallic catalysts enabled by machine learning. J. Mater. Chem. A. **5**, 24131–24138 (2017).
49. Gautreau, C., S. Sherry, S. Battista, A. Goldstein, and S. Stewart Enhancement motives moderate the relationship between high-arousal positive moods and drinking quantity: Evidence from a 22-day experience sampling study. Drug and Alcohol Review. **34**, 595–602 (2015).
50. O'Hara, R.E., S. Armeli, and H. Tennen College students' drinking motives and social-contextual factors: Comparing associations across levels of analysis. Psychology of Addictive Behaviors. **29**, 420–429 (2015).
51. Pleil, K.E., J.A. Rinker, E.G. Lowery-Gionta, C.M. Mazzone, N.M. McCall, A.M. Kendra, D.P. Olson, B.B. Lowell, K.A. Grant, T.E. Thiele, and T.L. Kash NPY signaling inhibits extended amygdala CRF neurons to suppress binge alcohol drinking. Nat. Neurosci. **18**, 545-+ (2015).
52. Koob, G.F. A role for brain stress systems in addiction. Neuron. **59**, 11–34 (2008).
53. Lindell, S.G., M.L. Schwandt, H. Sun, J.D. Sparenborg, K. Bjork, J.W. Kasckow, W.H. Sommer, D. Goldman, J.D. Higley, S.J. Suomi, M. Heilig, and C.S. Barr Functional NPY variation as a factor in

stress resilience and alcohol consumption in rhesus macaques. Arch. Gen. Psychiatry. **67**, 423–431 (2010).

Figures

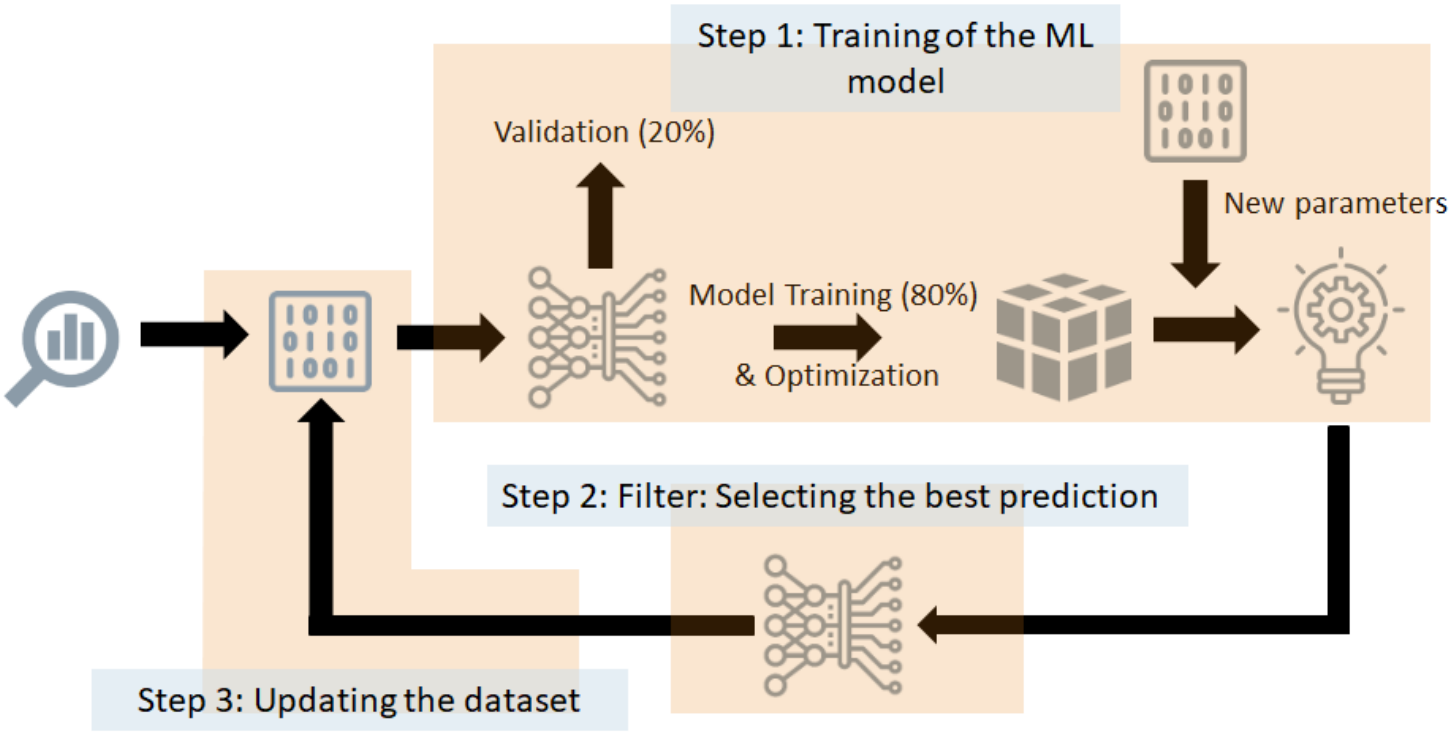
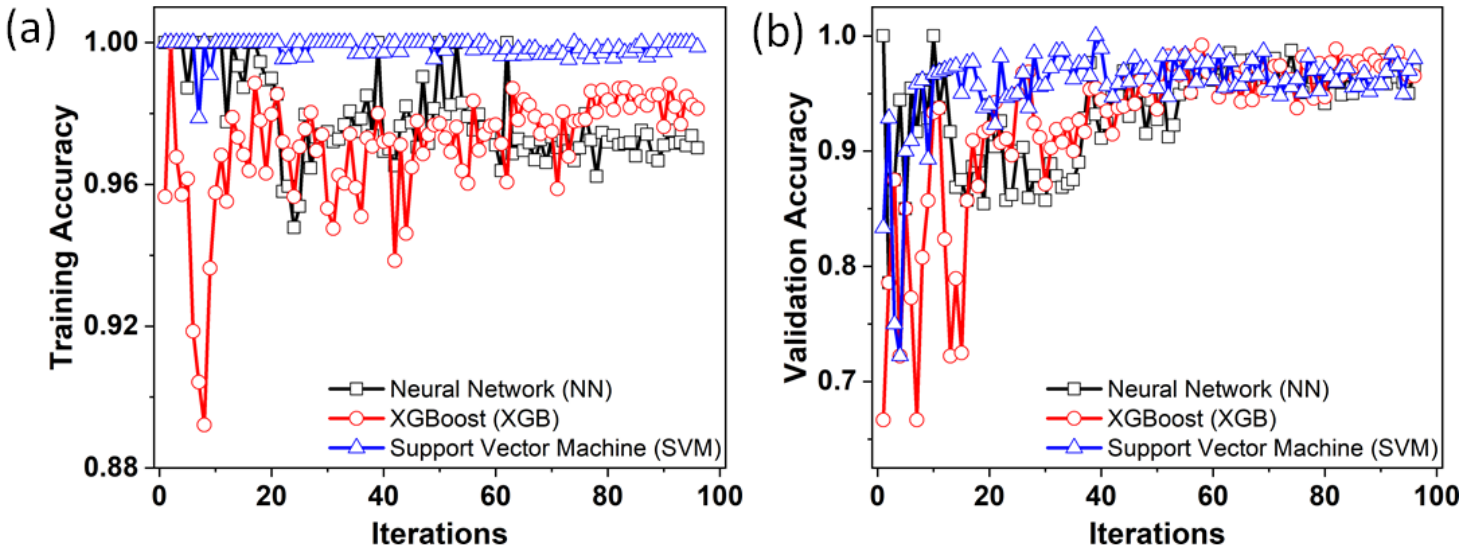


Figure 1

Schematic diagram of the SSG model. First, we trained a machine learning (ML) to predict the binge-drinking (BD) behavior on the basis of happiness (H), stress (S), attitude (A), subjective norm (SN), perceived control (PC), and intention (I). Subsequently, we generated new parameters within the constraint parameter spaces. Lastly, we employed a filtering strategy (FS) based on the statistical characteristic to ensure high-quality data augmentation. This cycle continued until the ~1000 dataset was achieved.



(c)

	Machine Learning Models					
	NN		XGB		SVM	
	Training Accuracy	Validation Accuracy	Training Accuracy	Validation Accuracy	Training Accuracy	Validation Accuracy
Average	0.98	0.94	0.97	0.93	1.00	0.96
Processor	Gforce RTX 2080 Ti		Gforce RTX 2080 Ti		Intel Core i7-9700	
Time	~4 min		~10 min		~ 1 min	

Figure 2

Evaluation of different ML models (NN, XGB, and SVM) based on the coefficient of determination (R^2). (a) Training accuracy, (b) validation accuracy, and (c) table of performance evaluation of the ML models. The data was split into 80% for the training set and 20% for the validation set on every iteration. At this stage, *FS* is not employed to reduce computational costs.

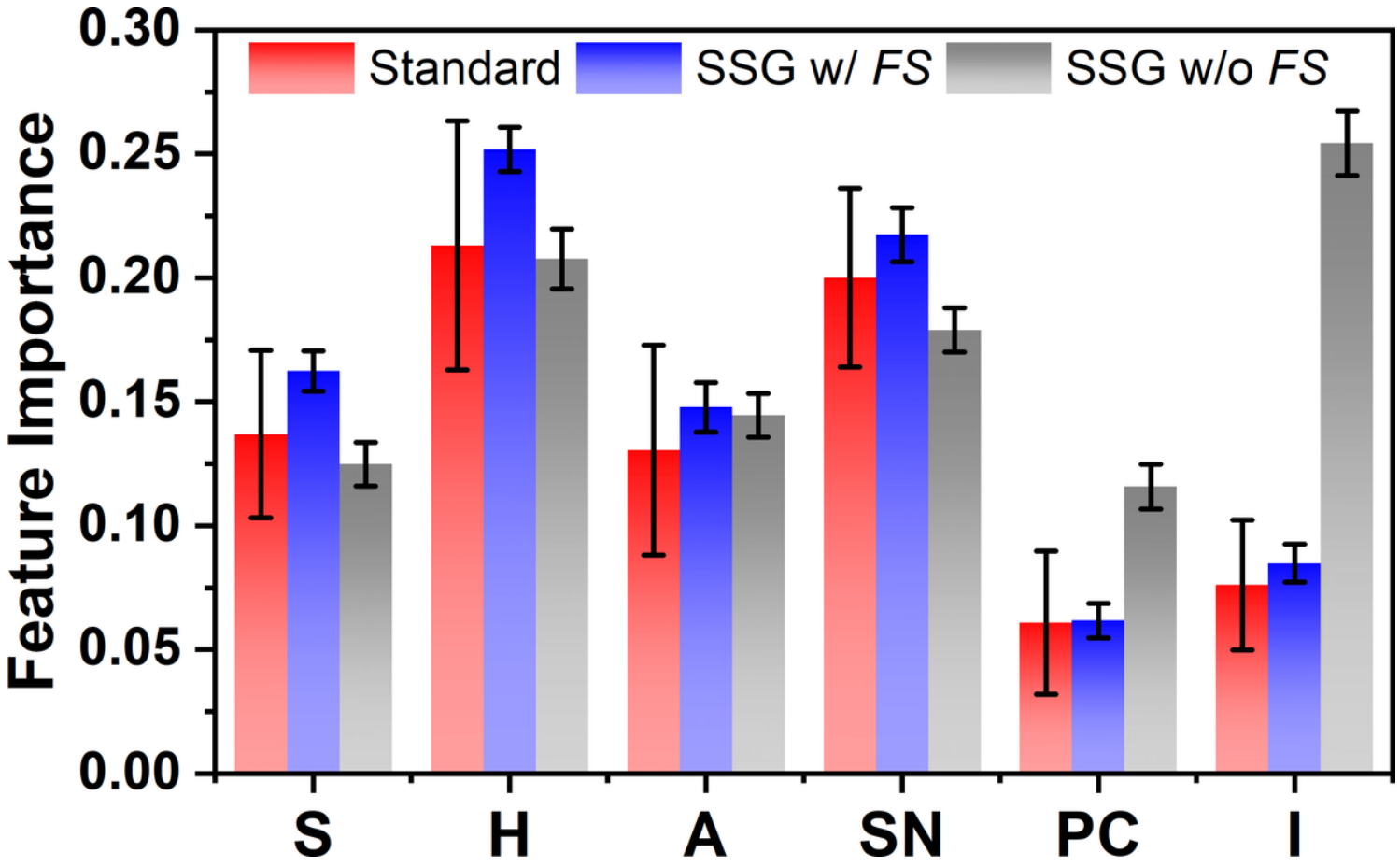


Figure 3

Visualization of the FI ranking before and after SSG operation. The FI of scarce and SSG with FS show a similar trend: happiness > subjective norm > stress > attitude > intention > perceived control. By contrast, SSG without FS implementation has a deviated trend: intention > happiness > subjective norm > attitude > stress > perceived control. The error bar represents the standard deviation from the average value of each feature.

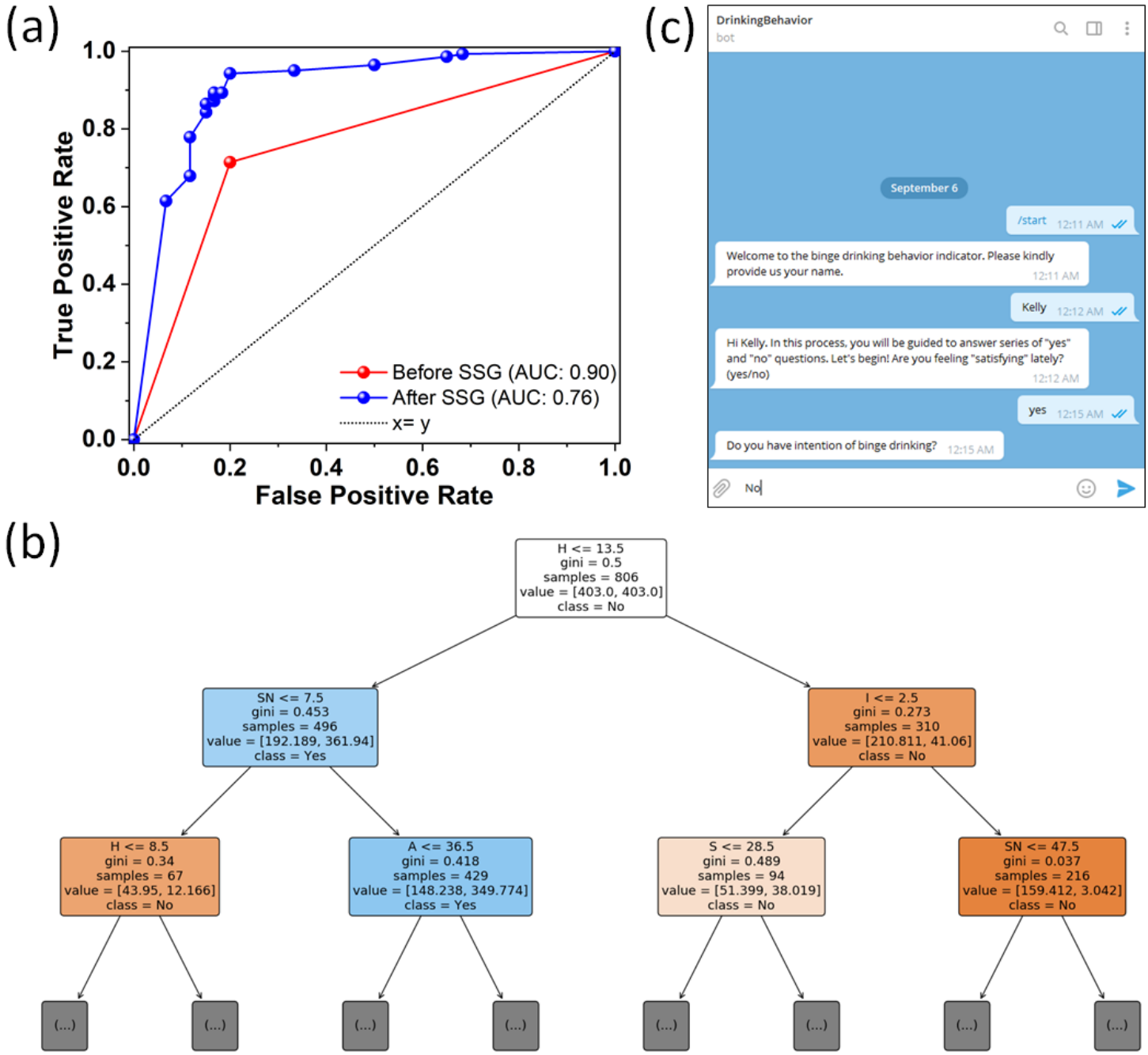


Figure 4

Decision tree (DT) analysis on the SSG augmented dataset. (a) Receiver operating characteristic (ROC) curves for the DT trained on scarce (standard) and SSG with FS. The dotted line and AUC represent the random guessing threshold and area under the curve. Good DT performance can be evaluated with the ROC curve above the guessing threshold and high AUC value. (b) Tree diagram of the SSG with the FS-augmented dataset based on the optimized DT model. (c) Telegram chatbot interface created based on the DT knowledge extraction of the SSG with the FS-augmented dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SILiow.docx