Machine Learning Prediction of Depression Risk from Behavioural and Self-Reported Data

Shuhan Yuan

{shuhan.yuan.24@ucl.ac.uk}

School of Management, University College London, London, United Kingdom

Abstract. The issue of depression has become a significant social concern in the world and there is need to use scalable methods to detect the risks at early onset. The paper explores how behavioural and demographic information predicts depression risk with the aid of several machine learning algorithms. Depression risk is identified using PHQ-9 screening tools, using NHANES survey data. The logistic regression, random forest, and gradient boosting are applied to assess the predictive performance of the behavioural-only and combined features. These findings demonstrate that behavioural variables by themselves have good predictive capacity as ROC-AUC is more than 0.80 in all models. Consistent but modest enhancements are given when the demographic variables are included. These results indicate that the measured lifestyle indicators can be used as informative variables to predict the presence of depression and outline the promise of machine learning strategies as scalable mental health screening of populations.

Keywords: Machine Learning; Depression Risk Prediction; Behavioural Data; Mental Health Analytics

1 Introduction

Depression has developed significant prevalence in the world in recent years and it is one of the primary public health concerns of countries [1]. Depression is linked with impairment of functions, low quality of life and high mortality risk [2]. It is thus important that they be detected early in the community to be intervened and prevented before their serious consequences. In the traditional method, the depression screening would be based on clinical structured instruments like Patient Health Questionnaire-9 (PHQ-9) which is a self-report questionnaire, used to gauge the level of symptoms experienced during the last two weeks [3]. Even though they are clinically tested, these tools involve active involvement and are normally implemented in the healthcare setting.

Meanwhile, mass behavioral data is becoming more and more available and population surveys and online. Such variables like sleep patterns, physical activity, alcohol intake and smoking habits might include predictive factors towards the state of mental health. The increasing access to behavioral data is an opportunity to establish scalable, data-driven approaches to depression risk identification beyond conventional screening based on symptomatic data.

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365164

Machine learning (ML) has been a widely used approach in mental health studies to establish patterns that can be used to detect a psychological condition [4]. Machine learning used in mental health analytics has various advantages regarding the sphere of diagnosis, treatment and support, research and clinical administration. The ML models can be trained using labelled data to understand the relationship between the predictors and the diagnostic results such that the probability of risk can be estimated. Some of the sources of data include electronic health records and neuroimaging data, survey responses, and social media activity.

The current literature has used machine learning methods on mental health prediction tasks [5]. They either stick to the symptom based measures or sophisticated models without contrasting the role of behavior measures against demographic features. Nevertheless, the current study fills these gaps by comparing the behavioral and combined feature sets (behavioral and demographic) in predicting the risk of depression among different machine learning algorithms. On the NHANES survey data and specifying the state of a depression risk through the PHQ-9 criteria, the results of the Logistic Regression, Random Forest, and Gradient Boosting models tested using F1-score and ROC-AUC at optimal thresholds were evaluated.

2 Data and methodology

2.1 Dataset description

The current study relies on the data of the National Health and Nutrition Examination Survey (NHANES), which is a large-scale, national survey that is used in the United States [6]. NHANES gathers the demographics, the behavioural and health aspects information via structured questionnaires, clinical examinations and laboratory tests. The dataset is their complete population-level healthcare source, which can be utilized to do predictive modelling.

The symptoms of depression are assessed with the help of the Patient Health Questionnaire-9 (PHQ-9), a self-reported assessment technique that is valid and screened. PHQ-9 scores vary between 0-27 and the higher the score the more the symptoms of depression are severe. In accordance with the clinical guidelines, people whose total PHQ-9 score exceeds 10 are considered to be at risk of depression level. This cutoff is common in epidemiological research [7]. Thus, binary outcome variable (DEPRESSION_RISK) is built.

Several modules of the NHAHES questionnaires are combined with the help of the unique participant identifier (SEQN). The analysis pool after merging and removing the missing value cases amounts to 8,276 people. The proportion of the population that meets the high-risk criteria (PHQ-9 $\geq$ 10) is about 9% which means that there is high imbalance between positive and negative cases. Therefore, the stratification of 80/20 ratio is observed to obtain 6,620 training observations and 1,656 testing observations.

2.2 Feature groups

In this paper, two sets of features are divided to examine the predictive value of the demographic information. The set of behavioral features has self-reported variables that can be associated with the risk level of depression, and includes alcohol intake, physical activity, sleep habits, and smoking habits. These behavioral indicators have been reliably observed in previous studies and are expected to contain predictive information beyond clinical symptom measures[8]. Isolation

of these features is aimed at evaluating the possibility of using routine data about lifestyle elements alone in giving useful predictor measures of depression risk.

The second configuration combined the behavioral characteristics with demographic characteristics like age, gender, ethnicity, education level, and household income. This study measures the predictive capability of the demographic information by comparing the results of behavioral-only and combined sets. This design works directly on the research question of whether behavioral data only are enough to predict depression risk or if demographic context is found to significantly improve predictive performance.

2.3 Preprocessing pipeline

All the preprocessing steps are carried out in line with scikit-learn pipeline framework to guarantee reproducibility and avoid information leaks and errors [9]. This was initiated by missing value treatment, scaling and encoding, and then train-test split. Since NHANES data is a survey-based dataset, missing values are also observed in various behavioral and demographic variables. Median imputation is the numerical imputation method applied to impute the numerical characteristics of data that was skewed and was not resistant to skewed distribution and outliers as is widely used in self-reports [10]. Categorical variables are coded in the most common category to ensure that they can be interpreted but are consistent across observations. After imputing, the categories of categorical variables are again encoded by way of a one-hot encoding scheme so that they could be incorporated into machine learning algorithms.

In order to test the generalized performance, the data is stratified using stratified sampling to create training (80) and testing (20) subsets, whereby the original class distribution in both subsets is maintained. The training data is used to conduct all model fitting, hyperparameter configuration and threshold optimization. The assorted evaluation measures are tested on the independent examination set.

2.4 Machine learning algorithms

This study uses three machine learning models to test the predictive tasks of behavioral and demographic variables in assessing the risk of depression: logistic regression, random forest, and gradient boosting. These models are linear and nonlinear modelling methods that are usually applied in health prediction tasks. One of the most common baseline classification algorithms is the logistic regression which attempts to predict the likelihood of a binary outcome by using a logistic regression model [11]. It is commonly used as a model of medical and epidemiological study.

Random forest refers to the ensemble approach to learning that involves the creation of a series of decision trees that have been built with the use of bootstrap samples of the training data [12]. Through the combination of predictions made by trees, random forest minimizes overfitting and nonlinear prediction relationships between predictors and outcomes. Also, another ensemble-based approach is known as gradient boosting which is additive; that is, each decision tree is constructed to address the errors in predictions of a previous model [13]. The reason why the iterative process of learning is useful is that it aids in comprehending intricate patterns in the data, and is much probable to deliver a robust prediction.

By combining these three algorithms one can make a comparison between an interpretable linear model and more relaxed ensemble techniques, offering an overall picture of the ability to predict based on various modelling assumptions.

3 Experiment results

3.1 Overall performance

Table 1 indicates the prediction performance of the three modelling algorithms in both configurations of features. In all three algorithms, there were the behavioural-only models that had a high predictive ability, and ROC-AUC was greater than 0.82. The results of logistic regression were ROC-AUC equal to 0.822 and PR-AUC equal to 0.303, whereas the results of both gradient boosting and random forest were 0.816 and 0.803, respectively. These findings demonstrate that behavioural variables only have significant predictive information in the classification of PHQ-9 based depression risk.

There was constant improvement in performance so far demographic factors were considered in all the modelling approaches. The highest ROC-AUC increase was observed in the logistic regression, and it is moving between 0.822 and 0.836. Afterwards, however, random forest and gradient boosting were going up to 0.828 and 0.832. The same pattern was observed in PR-AUC, and the regression model of the combination of logistics obtained the best value (0.325). The overall enhancement is relatively small, but still, it progresses in the same direction in all of the algorithms. This is an indication that demographic traits give incremental predictive value when compared to behavioral influences.

Table 1. Comparison of Models

Feature Set	Model	ROC-AUC	PR-AUC	Precision	Recall	F1-score
Behavioural	Logistic Regression	0.822	0.303	0.314	0.569	0.405
	Random Forest	0.803	0.243	0.296	0.601	0.397
	Gradient Boosting	0.816	0.291	0.262	0.647	0.373
Combined	Logistic Regression	0.836	0.325	0.374	0.464	0.414
	Random Forest	0.828	0.295	0.333	0.484	0.395
	Gradient Boosting	0.832	0.311	0.292	0.667	0.406

3.2 Comparison across models

Both feature arrangements and modelling algorithms have a number of constant trends. First, behavior-based models show good and consistent discrimination and ROC-AUC is more than 0.80 among the three algorithms. The discrepancies are not very high between the results of the logistic regression, random forest and gradient boosting. It means that comparable predictive capacity in linear and nonlinear approaches will appear in the cases of behavioral predictors only.

Secondly, demographic variables are included, which results in the overall enhancement of the performance of all modelling approaches. ROC-AUC and PR-AUC changes were increased between about 0.01 and 0.02. Although the extent of the improvement is weak, such improvements suggest that demographic data further provides predictive information over that of behavioral features.

Third, it is found that the complexity of the model does not affect overall ranking performance significantly. Logistic Regression realizes similar and in few cases slightly higher ROC-AUC and PR-AUC values compared to the ensemble methods. But a variation in class-specific measures comes out. Gradient Boosting has the best recall in the combined arrangement of features whereas Logistic Regression has the greatest accuracy. These differences merely represent the different trade offs in classification, but not markedly different levels of discrimination generally.

Overall, the findings suggest that there are no changes in predictive performance of the modelling strategies, linear and nonlinear approaches have incremental gains, and minimal difference between the two on the studied features space.

4 Discussion

4.1 Interpretation of findings

This paper has compared the predictive power of behavioural and demographic variables of PHQ-9 in defining depression risk in a variety of machine learning models. This study discovered a number of significant findings through the ML prediction. To begin with, behavioral variables alone performed well and stably on the discriminations in all three modelling techniques. This means that lifestyle predictors or variables like sleep patterns, smoking habits, alcohol use and physical exercise are useful at the population level and contain a significant amount of predictive information. The similarity of performance in a set of algorithms shows that behavioral cues are effective predictors, as opposed to the outcomes of a specific modelling strategy.

Additionally, the addition of the demographic variables led to consistent yet slight improvements in the predictive performance. Though the magnitude was small, there was an improvement in all the models. This trend indicates that demographic attributes bring out contextual refinements. Nevertheless, the demographics do not represent one of the major factors of classification performance. The behavioural predictors seem to explain a big percentage of the variance that is explainable by depression risk whereas demographic factors modify the probability estimates at the extremes.

In addition, the variations between linear and nonlinear models were not very high. The Logistic Regression showed similar levels of performance to the Random Forest and the Gradient Boosting in the overall ranking performance. The implication of this observation is that among the chosen feature space, predictor-depression risk relationships might be linear at the population scale. Although ensemble methods performed better than individual methods in some configurations, the failure to bring significant increase in ROC-AUC indicates that there is not much to be gained by making a model more complicated. Thus, these results indicate that behavioural data offer a significant predictive power and demographic variables offer consistent but poor improvement.

4.2 Limitations

This paper has several limitations that may be mentioned. To begin with, the clinical diagnosis assessment was not used to operationalize depression risk as the PHQ-9 criteria were used

instead. Although PHQ-9 is extensively valid, it is a self-reporting screening instrument and thus can create measurement error. Second, the data is not longitudinal and therefore does not provide the opportunity to predict or interpret causation. The models are estimations of the classification of risk based on association rather than the future occurrence of depressive disorder.

Thirdly, despite the weighting strategies and threshold optimization that guaranteed that the problem of class imbalance was mitigated, it is possible that residual imbalance affects precision-recall trade-offs. Also, the range of features available is restricted to self-reported behavioural and demographic variables included in NHANES. Any confounding factors or psychosocial variables that have not been observed can give an extra predictive signal. Lastly, the results are grounded in a U. S. sample population, which may not be applicable to other demographic and cultural settings without external confirmation.

4.2 Implications

These results have some implications for machine learning applications in mental health studies in future. First, the dominant superiority of behavioural-only models indicates that lifestyle measures that are regularly obtained could be useful as scalable depression risk screeners. This can be of possible interest to digital health systems and population-wide surveillance, in which clinical measures derived from symptoms are not available continuously.

Second, the small incremental change in the value of demographic factors suggests that behavioural cues might offer more immediate expressions of the risk of depression than the structural background features. This makes the relevance of the feature selection and domain-specific variable-building in the mental health prediction project evident.

Third, there is a small difference in the performance between linear and nonlinear models which leads to the belief that there are no proportional increases in predictive performance to increasing the complexity of algorithms in the structured survey data. This observation highlights how close attention should be paid to explain the complexity of the models used in practical mental health machine learning studies. Simpler models can offer similar utility in the context where model interpretability is relevant as well as transparency.

All in all, the paper adds to the methodological discourse on the feature configuration, model selection and imbalance conscious evaluation in the risk of depression prediction using population-level data.

5 Conclusions

This study examined how well behavioural and demographic data predict PHQ-9-defined This analysis examined how far behavioral and demographic variables could be employed to forecast PHQ-9 defined depression impending using various machine learning models. These results indicate that behavioral variables such as sleeping habits, smoking habits, alcoholism, and physical exercise offer a great predictive power. This makes high performance in discriminative performance in both linear and nonlinear algorithms. The stable behavioral-only models demonstrate that lifestyle-related indicators entail a significant signal related to the depressive risk at the population level.

Both the introduction of demographic variables had modest and consistent improvements across all modelling strategies. Although these advantages were not vast in scale, the findings indicate that demographic factors play an additive contextual optimization rather than the key factors in predictive power. Moreover, the fact that the performance difference between logistic regression and ensemble methods is narrow indicates that additional model complexity may not be an assurance of any improvement in the structured survey data.

Overall, the findings imply that behavioral indicators can provide scales and informative inputs to the classification of risks of depression, whereas demographic characteristics introduce an additional. The results of the research add to the ongoing debates on the topic of feature design, model choice, and the usefulness of machine learning in depression prediction. The study should be continued in the future by conducting longitudinal validation, additional behavioral signals and external replication that would help to determine the robustness and generalization of these results in other countries and situations.

References

[1] Lim, G.Y., Tam, W.W., Lu, Y., Ho, C.S., Zhang, M.W., Ho, R.C.: Prevalence of Depression in the Community from 30 Countries between 1994 and 2014. Scientific Reports. 8(1), p. 2861 (2018)

[2] Penninx, B.W., Milaneschi, Y., Lamers, F., Vogelzangs, N.: Understanding the somatic consequences of depression: biological mechanisms and the role of depression symptom profile. BMC Medicine. 11(1), p. 129 (2013)

[3] Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The PHQ-9. Journal of General Internal Medicine. 16(9), pp. 606-613 (2001)

[4] Shatte, A.B.R., Hutchinson, D.M., Teague, S.J.: Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine. 49(9), pp. 1426-1448 (2019)

[5] Iyortsuun, N.K., Kim, S.-H., Jhon, M., Yang, H.-J., Pant, S.: A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. Healthcare. 11(3), p. 285 (2023)

[6] U.S. Centers for Disease Control and Prevention, 'National Health and Nutrition Examination Survey', National Health and Nutrition Examination Survey. [Online]. Available: https://www.cdc.gov/nchs/nhanes/index.html

[7] Levis, B., Benedetti, A., Thombs, B.D.: Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ. 365, p. l1476 (2019)

[8] Cummins, N., Joshi, J., Dhall, A., Sethu, V., Goecke, R., Epps, J.: Diagnosis of depression by behavioural signals: a multimodal approach. In: Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge. pp. 11-20. Association for Computing Machinery, New York, NY, USA (2013)

[9] Kramer, O.: Scikit-Learn. In: Kramer, O. (ed.) Machine Learning for Evolution Strategies. pp. 45-53. Springer International Publishing, Cham (2016)

[10] Jadhav, A., Pramod, D., Ramanathan, K.: Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence. 33(10), pp. 913-933 (2019)

[11] LaValley, M.P.: Logistic Regression. Circulation. 117(18), pp. 2395-2399 (2008)

[12] Breiman, L.: Random Forests. Machine Learning. 45(1), pp. 5-32 (2001)

[13] Bentéjac, C., Csörgő, A., Martínez-Muñoz, G.: A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review. 54(3), pp. 1937-1967 (2021)