

Data mining and feature selection in machine learning analysis of cancer information seeking among U.S. adults

Ying Liu, Kesheng Wang

Submitted to: JMIR Medical Informatics
on: April 11, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	25

Preprint
JMIR Publications

Data mining and feature selection in machine learning analysis of cancer information seeking among U.S. adults

Ying Liu^{1*} PhD; Kesheng Wang^{2, 3*} PhD

¹ East Tennessee State University Johnson City US

² Cancer Survivorship Research Center, College of Nursing University of South Carolina Columbia US

³ Department of Epidemiology and Biostatistics, Arnold School of Public Health University of South Carolina Columbia US

*these authors contributed equally

Corresponding Author:

Ying Liu PhD

East Tennessee State University

P.O. Box 70259

Department of Biostatistics and Epidemiology, College of Public Health

Johnson City

US

Abstract

Background: Feature selection is essential in machine learning (ML) for identifying relevant variables. The Boruta algorithm and the least absolute shrinkage and selection operator (LASSO) are two widely used methods.

Objective: This study aimed to (1) compare feature selection methods (Boruta, LASSO, their combination, principal component analysis (PCA), and non-feature selection), and (2) develop ML tools to predict cancer information seeking among U.S. adults.

Methods: Data from 5505 individuals (2630 cancer information seekers and 2975 non-seekers) were selected from the 2022 Health Information National Trends Survey (HINTS 6). Four feature selection approaches and five ML tools (the support vector machines (SVMs) algorithms, logistic regression (LR), random forest (RF), k-nearest neighbor (KNN), and extreme gradient boosting (XGBoost)) were applied to develop ML models to predict cancer information seeking.

Results: The cancer information seeking prevalence was 47.2% (42.8% for males and 49.7% for females). The Boruta and LASSO selected 45 and 55 variables, respectively, with 36 in common. The PCA identified 21 uncorrelated factors. RF performed best, with similar AUCs for Boruta, LASSO, and no feature selection (?0.950) and accuracy (?0.860). Using PCA-selected variables yielded a slightly lower AUC (0.931) but comparable accuracy (0.853). Stepwise regression confirmed 21 of 36 selected key predictors, including personal/family cancer history, health information access, education, income, social media use, smoking, alcohol beliefs, and healthcare visits.

Conclusions: Feature selection effectively reduce dimensionality while retaining the predictive power. Boruta and LASSO performed comparably in terms of the selected variables. PCA-based selection uncorrelated s also proved useful. The key identified factors associated with cancer information seeking can guide future cancer intervention and prevention strategies.

(JMIR Preprints 11/04/2025:75862)

DOI: <https://doi.org/10.2196/preprints.75862>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/75862>, the full manuscript will be made available to all users.
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/75862>, the full manuscript will be made available to all users.



Original Manuscript

Data mining and feature selection in machine learning analysis of cancer information seeking among U.S. adults

Ying Liu^a, Kesheng Wang^{b,c}

^aDepartment of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN 37614, United States

^bCancer Survivorship Research Center, Department of Biobehavioral Health & Nursing Science, College of Nursing, University of South Carolina, Columbia, SC 29208, United States

^cDepartment of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, United States

Short title: Machine learning analysis of cancer information seeking

ABSTRACT

Background: Feature selection is essential in machine learning (ML) for identifying relevant variables. The Boruta algorithm and the least absolute shrinkage and selection operator (LASSO) are two widely used methods.

Objectives: This study aimed to (1) compare feature selection methods (Boruta, LASSO, their combination, principal component analysis (PCA), and non-feature selection), and (2) develop ML tools to predict cancer information seeking among U.S. adults.

Methods: Data from 5505 individuals (2630 cancer information seekers and 2975 non-seekers) were selected from the 2022 Health Information National Trends Survey (HINTS 6). Four feature selection approaches and five ML tools (the support vector machines (SVMs) algorithms, logistic regression (LR), random forest (RF), k-nearest neighbor (KNN), and extreme gradient boosting (XGBoost)) were applied to develop ML models to predict cancer information seeking.

Results: The cancer information seeking prevalence was 47.2% (42.8% for males and 49.7% for females). The Boruta and LASSO selected 45 and 55 variables, respectively, with 36 in common. The PCA identified 21 uncorrelated factors. RF performed best, with similar AUCs for Boruta, LASSO, and no feature selection (≈ 0.950) and accuracy (≈ 0.860). Using PCA-selected variables yielded a slightly lower AUC (0.931) but comparable accuracy (0.853). Stepwise regression confirmed 21 of 36 selected key predictors, including personal/family cancer history, health information access, education, income, social media use, smoking, alcohol beliefs, and healthcare visits.

Conclusions: Feature selection effectively reduce dimensionality while retaining the predictive power. Boruta

and LASSO performed comparably in terms of the selected variables. PCA-based selection uncorrelated s also proved useful. The key identified factors associated with cancer information seeking can guide future cancer intervention and prevention strategies.

Key words: Cancer information seeking; beliefs; knowledge; health behaviors; feature selection; PCA; machine learning

1. Introduction

With the development of science and technology, people seek health information from various sources, including online and offline approaches for different reasons, such as to understand personal or family members' health conditions, make decisions on health professional's recommendation [1]. According to U.S. National Cancer Institute report, approximately 40.5% of people will have cancer in their whole lifespan based on 2017-2019 data. Over two million people were diagnosed in the United States in 2024 [2]. Cancer patients and survivors desire information about their disease and its treatment, often actively seek cancer information to understand their condition, treatment approaches, and home care. their family member usually join them and seek more information. Healthy people may seek cancer information to enlarge their knowledge and improve their quality of life like changing their behaviors to healthy food, and have regular physical exams [3].

Machine learning (ML) and predictive analytics are commonly used in many areas, and they can transform data into useful insights for better understanding and faster decision-making. ML methods can address high dimensional data, model the etiological and clinical heterogeneity, and translate univariate variable findings into clinically useful multivariate decision support systems [4–8].

Feature selection is a critical step in ML not only to reduce the dimensionality of the feature space, but also to reveal the most relevant features without losing too much information [9–11]. Feature selection, as a preprocessing stage, is essentially the process of picking some informative and relevant features from a larger collection of features that produce a better characterization of patterns of multiple classes [12,13]. Several feature selection methods have been used in ML such as the least absolute shrinkage and selection operator (LASSO) [14–17] and the Boruta algorithm [14,15].

Principal component analysis (PCA) is an unsupervised feature reduction technique that explains the variance-covariance structure of a set of variables through linear combinations known as principal components or factors. Principal component analysis (PCA) has been used to reduce the dimensions but it is a not a feature selection method because all variables remained in each factor [13,18–20].

While numerous studies have investigated cancer information seeking behaviors [3,21,22];

however, few have employed ML approaches to identify the key factors that influence cancer information seeking behaviors. To address this gap, this study aimed to (1) compare four feature selection methods- Boruta, LASSO, combination of Boruta and LASSO, principal component analysis (PCA) and non-feature selection; and (2) develop ML tools to predict cancer information seeking among U.S. adults using the data from Health Information National Trends Survey (HINTS) 2022.

2. Methods

2.1. Sample

The data for this study was selected from the 2022 Health Information National Trends Survey (HINTS 6) included 6,252 respondents. The HINTS is a nationally representative survey administered by the National Cancer Institute (NCI) since 2003. The HINTS target adults aged 18 or older in the civilian non-institutionalized population of the United States (U.S.). The HINTS sponsored by the NCI, provide a unique opportunity to explore the characteristics of information seekers and non-seekers, as well as the content of information being sought by the public in a nationally representative sample. Data collection for HINTS 6 started on March 7, 2022, and concluded on November 8, 2022. The overall household response rate, based on the next-birthday method, was 28.1%. This study was exempt from the Institutional Review Board (IRB) approval since it involved secondary data analysis of a publicly accessible database.

2.2. Outcome variable

Subjects were considered to be cancer information seekers if they responded “yes” to the question “Have you ever looked for information about cancer from any source?”. Non-seekers were those if they responded “no” to the question.

2.3. Data processing of predictors

Demographic characteristics included gender, age group (18-49 years, 50-64 years, 65-74 years and 75 years or older), race, education, work full time (yes, no), income, and health insurance (yes, no). Race was recoded as Hispanic, non-Hispanic White, non-Hispanic Black or African American (AA), non-Hispanic Asian, and other. Education had four categories (less than high school, some college, bachelor's degree, and post-baccalaureate degree). The four categories of annual income were <\$19,999, \$20,000-49,999 and \$50,000-74,999 and 75,000+. Table 1 lists the demographic variables.

A total of 87 variables (including demographic factors and variables in alcohol and tobacco

use, health care, medical record, chronic diseases, beliefs about cancer, and health and nutrition etc.) were included in the initial analysis. Variables with missing value rate higher than 10% were removed. Individuals with missing values in outcome, age, gender and race were removed and then there are 5,505 individuals left. Race was generated for dummy variables. Other variables were binary or ordinal or continuous. The missing values in continuous variables were imputed by the mean and missing values in binary or ordinal variables were imputed by the mode. The full list of 73 predicting variables was listed in supplementary Table S1. Figure 1 shows an overview of the data curation and ML process.

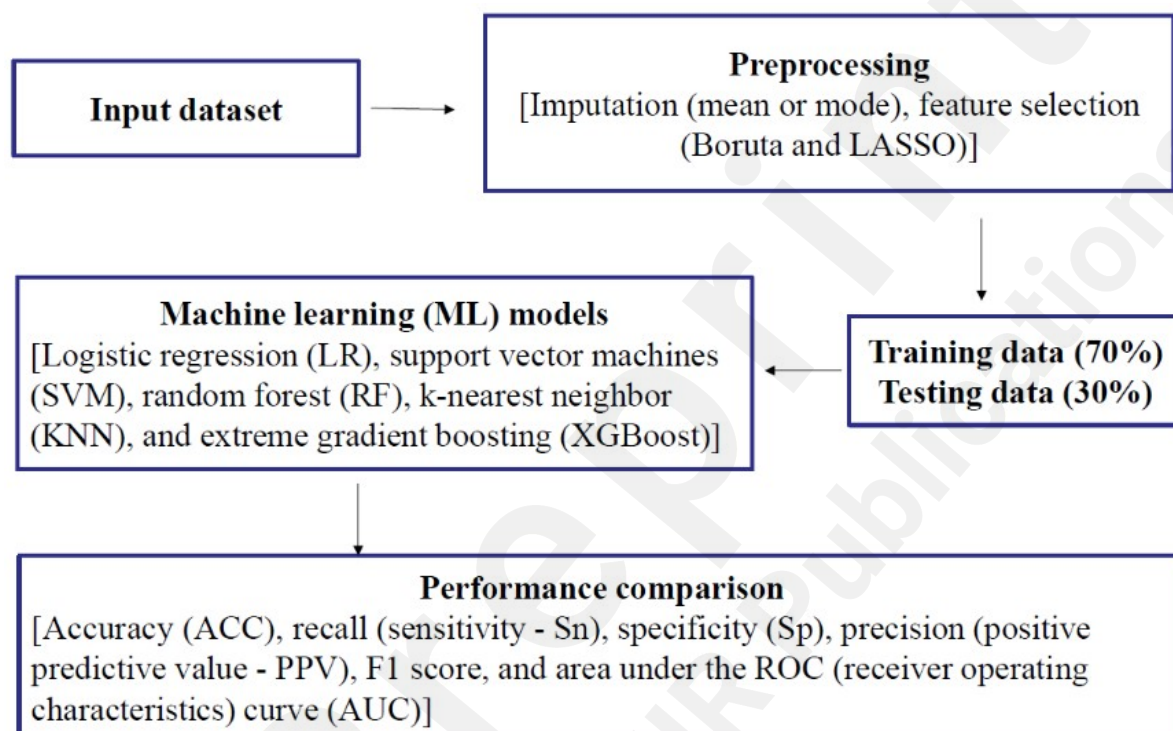


Figure 1. Overview of Data Curation and Machine Learning (ML) Workflow

2.4. Feature selection methods

2.4.1. Boruta

Boruta [23] and the least absolute shrinkage and selection operator (LASSO) [24] were used to select the variables associated with binary outcome. The Boruta algorithm, implemented in R via the “Boruta” package, automatically performs feature selection on a dataset using a random forest (RF) classifier [23].

2.4.2. LASSO

The LASSO feature selection was applied using “glmnet” package in R [24]. This method

regularizes the model by applying a penalty (λ), shrinking the regression coefficients and reducing some of them to zero. The feature selection phase occurs after the shrinkage, where non-zero values are selected as model parameters.

2.4.3. PCA

The PCA is an effective method to reduce the dimensionality of multivariate data. It captures the most variance in the dataset with a relatively small number of uncorrelated factors/PCs. Factor analysis was performed using PCA in SPSS. A Scree plot was used to visualize eigenvalues. The rotation method of Varimax (orthogonal) was used to produce different factor structures depending on the relationships between factors. A "rotated component matrix" in factor analysis is a table displaying the correlations between each variable in your dataset and the extracted factors after a rotation process has been applied, essentially making the factor structure easier to interpret by aligning the variables with their most relevant factors, often resulting in clearer groupings and higher loadings on specific factors compared to the unrotated solution. Large absolute values in a row of the matrix indicate that the corresponding variable has a strong relationship with that particular factor. Factors were selected based on eigenvalues greater than 1.0, and the corresponding factor scores were used for ML analyses. One variable with highest loading/correlation coefficients was chosen from each factor for further ML analyses.

2.5. Machine learning methods

Five ML algorithms were used, including Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), k-nearest neighbor (KNN), and Extreme Gradient Boosting (XGBoost). The SVM algorithm includes linear kernel, radial kernel, and polynomial kernel [25]. The RF algorithm randomly selects a subset of variables to construct multiple decision trees [26,27]. KNN, a simple ML algorithm based on a clustering algorithm with supervised learning, calculates the average of the numerical target of the K nearest neighbors. [28]. KNN is more suitable for low-dimensional data with a small number of input variables. XGBoost [29] is a supervised machine learning method for regression and classification tasks similar to R F classifier. The Caret package in R was used for LR, KNN, RF, and SVM [30]. The XGboost package was used for gradient boosting machine [31]. A 10-fold cross-validation approach was applied, and multiple parameters for each algorithm were optimized using a grid search.

2.6. Performance of machine learning

To evaluate the performance of feature selection methods, accuracy (ACC), recall (sensitivity - Sn), specificity (Sp), precision (positive predictive value - PPV), F1 score, and area under the ROC (receiver operating characteristics) curve (AUC). The R packages used included “caret”, “kernlab”, and “ROCR.”

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

$$Specificity = \frac{TN}{TN + FP} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$F1-Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

where *TP* is the number of true positives, *TN* is the number of true negatives, *FP* is the number of false positives, and *FN* is the number of false negatives. F1-Score is a harmonic mean that combines both recall and precision.

2.7. Statistical analysis

The categorical variables were presented in their raw values along with the proportions for categorical variables. The Chi-square test was used to examine the associations of categorical variables with cancer information seeking across demographic variables. Stepwise logistic regression analysis was performed to estimate the odds ratios (ORs) and their 95% confidence intervals (CIs) for the associations of potential factors with outcome. All statistical analyses were performed using SPSS software, version 29.

3. Results

3.1. Prevalence of cancer information seeking

Among 5,505 adult respondents, there were 2,630 cancer information seekers and 2,975 non-seekers (Table 1). The overall prevalence was 47.2% (42.8% for males and 48.7% for females). The prevalence increased with age (46.3%, 49.2%, 50.5% for age groups 18-49, 50-64, and 65-74, respectively). The age group of 75+ has a lower prevalence (44.0%). The prevalence was higher in those with higher education and higher income.

3.2. Feature selection

The Boruta algorithm selected 45 variables and LASSO selected 55 variables related to cancer information seeking (Supplementary table S1). Of these, 36 variables were identified by both methods. The PCA identified 21 uncorrelated factors with eigenvalues greater than 1 (Figure 2 and supplementary table S2). From each factor, we selected the variable with the highest loading/correlation (Supplementary table S3).

3.3. Machine learning performance

The performance statistics are summarized in Table 2. Based on accuracy and AUC, RF emerged as the best ML model with highest AUC across all 6 feature selection methods, and highest accuracy in five out of six feature selection methods (Table 2). Based on RF, Boruta, LASSO, combination of Boruta and LASSO, and non-feature selection yielded the almost same AUC (0.949, 0.950, 0.949 and 0.950, respectively) and similar accuracy (0.860, 0.876, 0.855 and 0.857, respectively). The second-best ML model is XGboost, Boruta, LASSO, combination of Boruta and LASSO, and non-feature selection revealed the almost same AUC (0.910, 0.912, 0.924 and 0.931, respectively) and similar accuracy (0.844, 0.850, 0.857 and 0.852, respectively). When selecting one variable from each of 21 factors with highest loading with the factor, the RF model showed a slightly lower AUC (0.931) than the model using 36 variables from both Boruta and LASSO but maintained similar accuracy (0.853).

Plot of mean decrease accuracy and mean decrease Gini values using random forest algorithm for 36 variables selected by LASSO and Boruta was illustrated in Figure 3 and those for the 21 variables from each factor were illustrated in Figure 4. (Figure 4). Plot of mean decrease accuracy and mean decrease Gini values using RF algorithm, XGBoost, GBM, LR models and the 21 factors are illustrated in supplementary figures s1-s4. The AUC curves in the validation data for RF model, based on LASSO, Boruta, 35 variables selected by both LASSO and Boruta, all 73 variables, 21 factor scores and the 21 variables from each factor were illustrated in Figure 5.

3.4. Stepwise logistic regression analysis

Stepwise logistic regression of 36 variables identified by both Boruta and LASSO and further confirmed 21 variables were associated with cancer information seeking (Table 3). The top variables included having cancer, a family history of cancer, concerns about developing cancer, knowledge of genetic test and HPV, access to health records, higher income, higher education, frequent social-

media video consumption, belief in alcohol increasing cancer and older age, smoking status, perception of the harm of e-cigarette use, binge drinking, and frequency of healthcare visits.

4. Discussion

To the best of our knowledge, this study represents the inaugural attempt to focus on feature selection and the development and validation ML tools to predict cancer information seeking behavior using the latest HINTS 2022 data. Both Boruta and LASSO identified 36 identical variables. Based on the 35 selected variables, the RF achieved the highest accuracy. A stepwise logistic regression model further confirmed that 21 of the 36 variables were significantly associated with cancer information seeking.

Feature selection is a critical step in ML to reduce the dimensionality while retaining the most relevant features without excessive information loss [9–11]. Previous studies have used several feature selection methods including LASSO, Boruta, and RF [14–17]. In the present study, we applied Boruta and LASSO methods and selected 36 overlapping variables (Table 2). Boruta algorithm is a feature selection method based on random forest and considers important variables but does not account for collinearity while selecting variables, whereas LASSO regression is useful when multicollinearity exists in the model [23,32,33]. The LASSO can be used as a variable selection method because numerous beta coefficients that are not strongly associated with the outcome are decreased to zero, which is equivalent to removing those variables from the model. However, the disadvantage of LASSO is that it assumes a more restrictive set of assumptions than the random forest. For example, that the LASSO is a linear model, so anything that matters for your outcome that isn't linear in the parameters under estimation is at risk of getting eliminated.

ML represents a powerful tool that could advance tobacco control research and policy decision-making [34]. The present study showed that the RF model was the best classifier and XGBoost was the second-best model (Table 2).

Cancer patient often seek information regarding their diagnosis, treatment options, treatment costs, potential side effects, and the implications for daily life and survival. While physicians remain a primary source of such information, patients frequently turn to additional resources such as the internet and books to supplement their understanding [21]. A previous study has identified that various factors influence cancer information-seeking behaviors, including gender, education levels, income, and cancer type [3]. In the present study, machine learning techniques were employed to identify key predictors of cancer information seeking behavior. The most influential factors included

personal or family history of cancer, access to health information, education, income, social media use, smoking status, beliefs about alcohol, and frequency of healthcare visits.

4.1. Strengths and limitations

This study has several notable strengths. First, this study used the most recent HINTS (2022) data to examine the prevalence of cancer information seeking. The HINTS data provides unparalleled insights into health information seeking behaviors, social media use, beliefs about alcohol and cancer etc. Second, we performed feature selection using two widely used methods, LASSO and Boruta, to identify common variables across both methods. Third, we inferred PC scores (factor scores) and then used weighted logistic regression analyses to estimate the associations of potential factors and PC scores with CRC screening. Fourth, we compared 5 ML algorithms and found the RF model demonstrated outstanding classification performance in predicting cancer information seeking. Moreover, we used stepwise logistic regression analysis to validate the results from ML techniques.

Despite these strengths, our analysis has some limitations. Firstly, because the HINTS data are cross-sectional, we could only identify correlations rather than causal relationships. Future research could address this limitation by applying a rigorous quasi-experimental method to longitudinal datasets. Secondly, since information from participants was self-reported, our study may be subject to recall bias as well as social desirability bias. Furthermore, the data were collected in 2022, the COVID-19 pandemic may influence both data collection and results.

5. Conclusion

The present study provided the updated prevalence of cancer information seeking among U.S. adults. Furthermore, we performed feature selection and compared five ML algorithms for classifying cancer information seeking, and identified that the RF was the best performer with the highest accuracy and AUC. Based on the stepwise regression model, 21 out of 36 selected variables were significantly associated with cancer information seeking. Moreover, PCA proved useful for data mining to reduce the indicators in complex survey data and aid in feature selection. Our findings may benefit researchers, policymakers, and health care providers by increasing public awareness and supporting targeted education on cancer information seeking.

Funding

No funding source is given for the present paper.

CRediT authorship contribution statement

Ying Liu: Writing – original draft, Software, Methodology, Formal analysis, Data curation.
Kesheng Wang: Writing – original, review & editing, Software, Methodology, Formal analysis, Conceptualization.

Ethical Considerations

This study used secondary data. The HINTS data meet strict ethical standards and have obtained ethics approval. Informed consent was obtained from all participants, and all methods were carried out in accordance with relevant guidelines and regulations.

Data availability statement

The data that support the findings of this study are openly available in [<https://hints.cancer.gov/data/default.aspx>] at <https://hints.cancer.gov/>.

Acknowledgements

The authors would like to thank the NCI for providing the data from the 2022 Health Information National Trends Survey.

Appendix A. Supplementary data

Table S1. Feature selection based on LASSO and Boruta.

Table S2. Eigenvalues based on PCA.

Table S3. Rotated component/loading of variables for 21 factors.

Figures S1-S4. Plot of mean decrease accuracy (left panel) and mean decrease gini (right panel) values using random forest algorithm and LASSO, Boruta, all 73 variables and 21 factors.

References

1. Weaver JB, Mays D, Weaver SS, Hopkins GL, Eroğlu D, Bernhardt JM. Health Information–Seeking Behaviors, Health Indicators, and Health Risks. *Am J Public Health* 2010 Aug;100(8):1520–1525. doi: 10.2105/AJPH.2009.180521
2. National Cancer Institute. Cancer Statistics. 2024. Available from: <https://www.cancer.gov/about-cancer/understanding/statistics> [accessed Mar 20, 2025]
3. Roach A, Lykins E, Gochett C, Brechting E, Graue L, Andrykowski M. Differences in Cancer Information-Seeking Behavior, Preferences, and Awareness Between Cancer Survivors and

- Healthy Controls: A National, Population-Based Survey. *J of Cancer Educ* 2009 Jan;24(1):73–79. doi: 10.1080/08858190802664784
4. Hurtado M, Siefkas A, Attwood MM, Iqbal Z, Hoffman J. Machine Learning Applications and Advancements in Alcohol Use Disorder: A Systematic Review. *Addiction Medicine*; 2022 Jun. doi: 10.1101/2022.06.06.22276057
 5. Kufel J, Bargieł-Łączek K, Kocot S, Koźlik M, Bartnikowska W, Janik M, Czogalik Ł, Dudek P, Magiera M, Lis A, Paszkiewicz I, Nawrat Z, Cebula M, Gruszczyńska K. What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. *Diagnostics* 2023 Aug 3;13(15):2582. doi: 10.3390/diagnostics13152582
 6. Theodosiou AA, Read RC. Artificial intelligence, machine learning and deep learning: Potential resources for the infection clinician. *Journal of Infection* 2023 Oct;87(4):287–294. doi: 10.1016/j.jinf.2023.07.006
 7. Kumari J, Kumar E, Kumar D. A Structured Analysis to study the Role of Machine Learning and Deep Learning in The Healthcare Sector with Big Data Analytics. *Arch Computat Methods Eng* 2023 Jul;30(6):3673–3701. doi: 10.1007/s11831-023-09915-y
 8. Shastry KA, Vijayakumar V, V MKM, B A M, B N C. Deep Learning Techniques for the Effective Prediction of Alzheimer’s Disease: A Comprehensive Review. *Healthcare (Basel)* 2022 Sep 23;10(10):1842. PMID:36292289
 9. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing* 2018 Jul;300:70–79. doi: 10.1016/j.neucom.2017.11.077
 10. Awan SE, Bennamoun M, Sohel F, Sanfilippo FM, Chow BJ, Dwivedi G. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. Cheungpasitporn W, editor. *PLoS ONE* 2019 Jun 26;14(6):e0218760. doi: 10.1371/journal.pone.0218760
 11. Cömert Z, Şengür A, Budak Ü, Kocamaz AF. Prediction of intrapartum fetal hypoxia considering feature selection algorithms and machine learning models. *Health Inf Sci Syst* 2019 Dec;7(1):17. doi: 10.1007/s13755-019-0079-z
 12. Chen X, Kopsaftopoulos F, Wu Q, Ren H, Chang F-K. Flight State Identification of a Self-Sensing Wing via an Improved Feature Selection Method and Machine Learning Approaches. *Sensors* 2018 Apr 29;18(5):1379. doi: 10.3390/s18051379
 13. Raihan-Al-Masud Md, Mondal MRH. Data-driven diagnosis of spinal abnormalities using feature selection and machine learning algorithms. Pérez MA, editor. *PLoS ONE* 2020 Feb 6;15(2):e0228422. doi: 10.1371/journal.pone.0228422
 14. Atuegwu NC, Litt MD, Krishnan-Sarin S, Laubenbacher RC, Perez MF, Mortensen EM. E-Cigarette Use in Young Adult Never Cigarette Smokers with Disabilities: Results from the Behavioral Risk Factor Surveillance System Survey. *IJERPH* 2021 May 20;18(10):5476. doi: 10.3390/ijerph18105476
 15. Atuegwu NC, Oncken C, Laubenbacher RC, Perez MF, Mortensen EM. Factors Associated with E-Cigarette Use in U.S. Young Adult Never Smokers of Conventional Cigarettes: A Machine

Learning Approach. IJERPH 2020 Oct 5;17(19):7271. doi: 10.3390/ijerph17197271

16. Choi J, Jung H-T, Ferrell A, Woo S, Haddad L. Machine Learning-Based Nicotine Addiction Prediction Models for Youth E-Cigarette and Waterpipe (Hookah) Users. JCM 2021 Mar 2;10(5):972. doi: 10.3390/jcm10050972
17. Fu R, Shi J, Chaiton M, Leventhal AM, Unger JB, Barrington-Trimis JL. A Machine Learning Approach to Identify Predictors of Frequent Vaping and Vulnerable Californian Youth Subgroups. Nicotine & Tobacco Research 2022 Jun 15;24(7):1028–1036. doi: 10.1093/ntr/ntab257
18. Castillo-Barnes D, Su L, Ramírez J, Salas-Gonzalez D, Martinez-Murcia FJ, Illan IA, Segovia F, Ortiz A, Cruchaga C, Farlow MR, Xiong C, Graff-Radford NR, Schofield PR, Masters CL, Salloway S, Jucker M, Mori H, Levin J, Gorriz JM, (Dian) DIAN. Autosomal dominantly inherited alzheimer disease: Analysis of genetic subgroups by machine learning. Information Fusion 2020 Jun;58:153–167. doi: 10.1016/j.inffus.2020.01.001
19. Chen C, Qin Y, Cheng J, Gao F, Zhou X. Texture Analysis of Fat-Suppressed T2-Weighted Magnetic Resonance Imaging and Use of Machine Learning to Discriminate Nasal and Paranasal Sinus Small Round Malignant Cell Tumors. Front Oncol 2021 Dec 13;11:701289. doi: 10.3389/fonc.2021.701289
20. Grazioli G, Martin RW, Butts CT. Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods. Front Mol Biosci 2019 Jun 12;6:42. doi: 10.3389/fmolb.2019.00042
21. Nagler RH, Gray SW, Romantan A, Kelly BJ, DeMichele A, Armstrong K, Schwartz JS, Hornik RC. Differences in information seeking among breast, prostate, and colorectal cancer patients: Results from a population-based survey. Patient Education and Counseling 2010 Dec;81:S54–S62. doi: 10.1016/j.pec.2010.09.010
22. Rexhepi H, Huvila I, Åhlfeldt R-M, Cajander Å. Cancer patients' information seeking behavior related to online electronic healthcare records. Health Informatics J 2021 Jul;27(3):14604582211024708. doi: 10.1177/14604582211024708
23. Kursa MB, Rudnicki WR. Feature Selection with the **Boruta** Package. J Stat Soft 2010;36(11). doi: 10.18637/jss.v036.i11
24. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Soft 2010;33(1). doi: 10.18637/jss.v033.i01
25. Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995 Sep;20(3):273–297. doi: 10.1007/BF00994018
26. Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J. A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. CATENA 2017 Apr;151:147–160. doi: 10.1016/j.catena.2016.11.032
27. Kesler SR, Rao A, Blayney DW, Oakley-Girvan IA, Karuturi M, Palesh O. Predicting Long-Term Cognitive Outcome Following Breast Cancer with Pre-Treatment Resting State fMRI and

- Random Forest Machine Learning. *Front Hum Neurosci* 2017 Nov 15;11:555. doi: 10.3389/fnhum.2017.00555
28. Zhang Z. Introduction to machine learning: k-nearest neighbors. *Ann Transl Med* 2016 Jun;4(11):218–218. doi: 10.21037/atm.2016.03.37
 29. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco California USA: ACM; 2016. p. 785–794. doi: 10.1145/2939672.2939785*
 30. Kuhn M. Building Predictive Models in *R* Using the **caret** Package. *J Stat Soft* 2008;28(5). doi: 10.18637/jss.v028.i05
 31. XGboost: Extreme Gradient Boosting. *R Package Version 0.4-2* 1, 1–4.
 32. Kaneko H. Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables. *Heliyon* 2021 Jun;7(6):e07356. doi: 10.1016/j.heliyon.2021.e07356
 33. Wang S, Nan B, Rosset S, Zhu J. Random lasso. *Ann Appl Stat* 2011 Mar 1;5(1). doi: 10.1214/10-AOAS377
 34. Fu R, Kundu A, Mitsakakis N, Elton-Marshall T, Wang W, Hill S, Bondy SJ, Hamilton H, Selby P, Schwartz R, Chaiton MO. Machine learning applications in tobacco research: a scoping review. *Tob Control* 2023 Jan;32(1):99–109. doi: 10.1136/tobaccocontrol-2020-056438

Table 1. Prevalence of cancer information seeking across demographic factor (%)

Variable	Total (N)	Seeking	Prevalence (%) 95% CI	p-value
Gender				
Male	2307	987	42.8	<0.001
Female	3535	1758	49.7	
Age group				
18-49 years	1935	895	46.3	0.012
50-64 years	1607	791	49.2	
65-74 years	1234	623	50.5	
75+ years	729	321	44.0	

Race				
Non-Hispanic White	3175	1738	54.7	<0.001
Non-Hispanic African American	878	314	35.8	
Hispanic	984	371	37.7	
Non-Hispanic Asian	286	120	42.0	
Other	182	875	47.8	
Education				
Less than High School	1300	375	28.8	<0.001
Some College	1574	721	45.8	
Bachelor's Degree	1550	843	54.4	
Post-Baccalaureate Degree	1070	687	64.2	
Income				
<19,999	904	271	30.0	<0.001
20,000-49,999	1431	569	39.8	
50,000-74,999	946	488	51.6	
75,000+	2224	1302	58.5	
Overall	5505	2630	47.2	

Abbreviations: CI: Confidence interval. p-value is based on χ^2 test.

Table 2. Machine learning and comparison of performance

Model	Variation	Features	Accuracy	Sensitivity (Recall)	Specificity	Precision	F1-Score	AUC
SVM	Linear kernel	All features	0.703	0.696	0.710	0.710	0.703	0.786
		LASSO	0.698	0.694	0.703	0.704	0.699	0.782
		Boruta	0.687	0.681	0.694	0.694	0.687	0.781
		LASSO and Boruta	0.686	0.678	0.695	0.694	0.686	0.780
		PCA score	0.694	0.670	0.717	0.707	0.688	0.768
		PCA based - highest loading variable	0.662	0.638	0.686	0.674	0.656	0.717
	RBF kernel	All features	0.841	0.881	0.800	0.818	0.848	0.944

		LASSO	0.845	0.869	0.820	0.831	0.850	0.936
		Boruta	0.776	0.595	0.959	0.938	0.728	0.932
		LASSO and Boruta	0.827	0.921	0.732	0.778	0.843	0.937
		PCA score	0.835	0.883	0.785	0.807	0.843	0.932
		PCA based - highest loading variable	0.823	0.908	0.736	0.778	0.838	0.920
LR	Logistic regression	All features	0.699	0.696	0.703	0.704	0.700	0.786
		LASSO	0.698	0.679	0.717	0.710	0.694	0.785
		Boruta	0.693	0.694	0.694	0.698	0.696	0.783
		LASSO and Boruta	0.694	0.685	0.703	0.701	0.693	0.782
		PCA score	0.682	0.656	0.708	0.696	0.675	0.770
RF	Random forest	PCA based - highest loading variable	0.660	0.641	0.679	0.670	0.655	0.718
		All features	0.857	0.856	0.859	0.861	0.858	0.950
		LASSO	0.876	0.878	0.873	0.876	0.877	0.950
		Boruta	0.860	0.867	0.853	0.857	0.862	0.949
		LASSO and Boruta	0.855	0.858	0.851	0.855	0.856	0.949
KNN	k-nearest neighbors	PCA score	0.850	0.872	0.828	0.837	0.834	0.944
		PCA based - highest loading variable	0.853	0.858	0.848	0.852	0.855	0.931
		All features	0.813	0.789	0.837	0.831	0.809	0.814
		LASSO	0.815	0.798	0.831	0.828	0.813	0.815
		Boruta	0.825	0.793	0.857	0.849	0.820	0.818
XGBoost	Extreme gradient boosting	LASSO and Boruta	0.835	0.811	0.859	0.854	0.832	0.815
		PCA score	0.806	0.775	0.857	0.829	0.801	0.808
		PCA based - highest loading variable	0.820	0.804	0.837	0.834	0.819	0.806
		All features	0.852	0.858	0.846	0.850	0.854	0.931
		LASSO	0.850	0.850	0.844	0.848	0.849	0.912
		Boruta	0.844	0.847	0.840	0.844	0.845	0.910
		LASSO and Boruta	0.857	0.854	0.861	0.862	0.858	0.924
		PCA score	0.856	0.869	0.842	0.849	0.859	0.938
		PCA based - highest loading variable	0.835	0.823	0.846	0.845	0.834	0.908

Abbreviations: SVM: Support vector machine; RBF: Radial basis function; AUC: Area under the ROC (receiver operating characteristics) curve.

Table 3. Stepwise logistic regression analyses of 36 variables

Variable	OR (95% CI)	p-value
Gender (ref=male)		
Female	1.23 (1.08-1.41)	0.002
Education (ref= less than high school)		
Some College	1.40 (1.17-1.68)	<0.001
Bachelor's Degree	1.48 (1.22-1.80)	<0.001
Post-Baccalaureate Degree	1.84 (1.48-2.29)	<0.001

Income (ref=<19,999)		
20,000-49,999	1.18 (0.96-1.46)	0.115
50,000-74,999	1.63 (1.29-2.06)	<0.001
75,000+	1.74 (1.39-2.17)	<0.001
Race (ref=White)		
African American	0.62 (0.51-0.74)	<0.001
Hispanic	0.80 (0.66-0.95)	0.013
Asian	0.84 (0.63-1.12)	0.221
Trust charities (1-4, 4= a lot)	1.10 (1.02-1.19)	0.012
Alcohol_intent (1-4, 4= drink more alcohol)	1.10 (1.02-1.18)	0.010
Alcohol_Increase_Cancer (1-4, 4= a lot)	1.24 (1.14-1.35)	<0.001
Meat_Increase_Cancer (1-4, 4= a lot)	1.11 (1.04-1.18)	<0.001
Confident_Internet_Health Resource (1-5, 5=not confident at all)	1.10 (1.02-1.18)	0.013
Cancer (ref=no)		<0.001
Yes	2.84 (2.33-3.45)	<0.001
Family had cancer (ref=no)		
Yes	0.80 (0.62-1.04)	0.091
Not sure	1.70 (1.35-2.15)	<0.001
Heard_genetic_test (Ref=no)		
Yes	1.65 (1.38-1.98)	<0.001
Heard_HPV(Ref=no)		
Yes	1.43 (1.24-1.66)	<0.001
Doctor told colorectal Cancer Tests (ref=never)		
No	1.08 (0.90-1.30)	0.416
Yes	1.22(1.04-1.43)	0.017
Prevention cancer not possible (1-4, 4=strongly agree)	0.86 (0.79-0.92)	<0.001
Worry_get_cancer (1-5, 5=extremely)	1.24 (1.17-1.31)	<0.001
Frequency_go_provider last year (1-6, 6=10 or more times)	1.05 (1.01-1.08)	0.015
Offer_access online medical record (ref= do not know)		
No	1.13 (0.84-1.15)	0.416
Yes	1.60 (1.23-2.09)	<0.001
Social_media_misleading health infor (1-5, 5= a lot)	1.08 (1.03-1.13)	0.002
Social_media_watch_video (1-5, 5=almost everyday)	1.24 (1.16-1.32)	<0.001
COVID19 delay cancer_screening (1-5, 5= a lot)	1.08 (1.02-1.15)	0.012

Abbreviations: OR = Adjusted odds ratio; CI = Confidence interval.

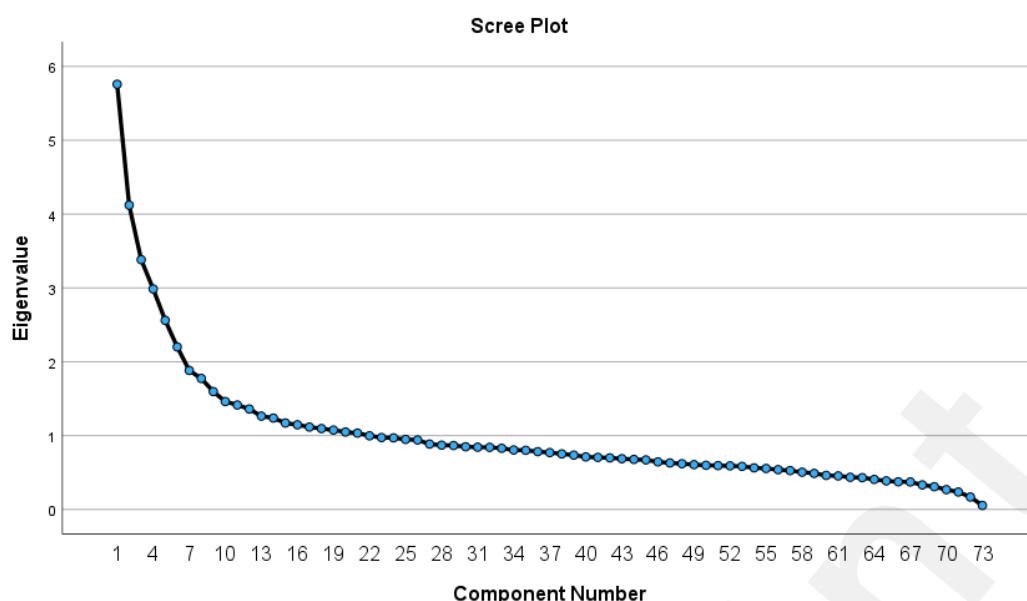


Figure 2. Scree plot showing 21 factors with eigenvalue higher than 1

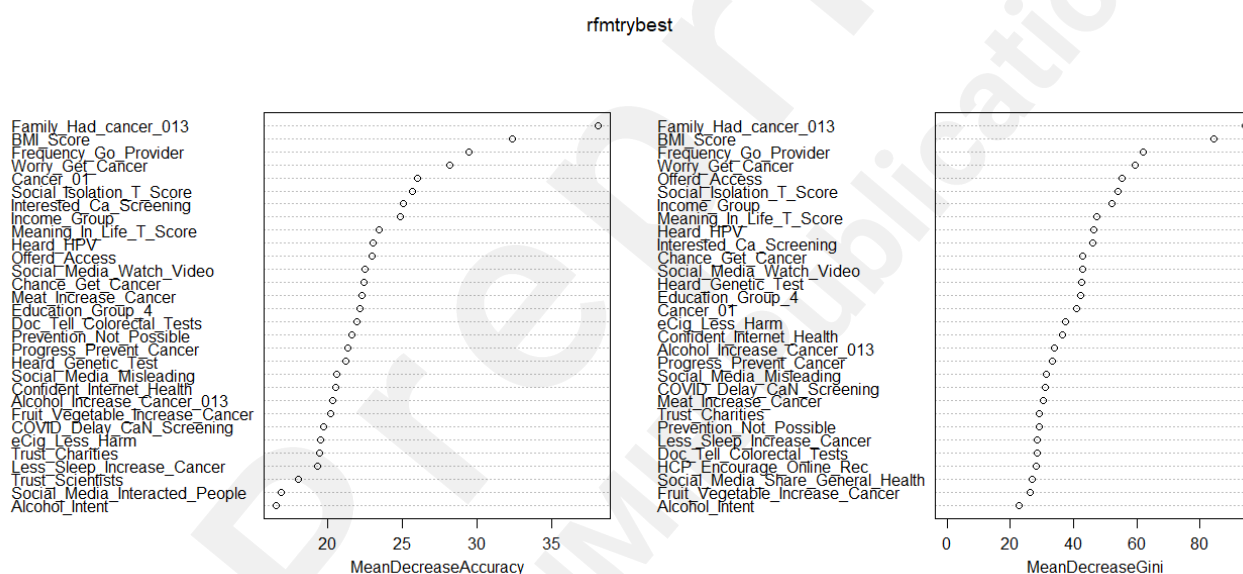


Figure 3. Plot of mean decrease accuracy (left panel) and mean decrease gini (right panel) values using random forest algorithm and 36 variables selected by LASSO and Boruta

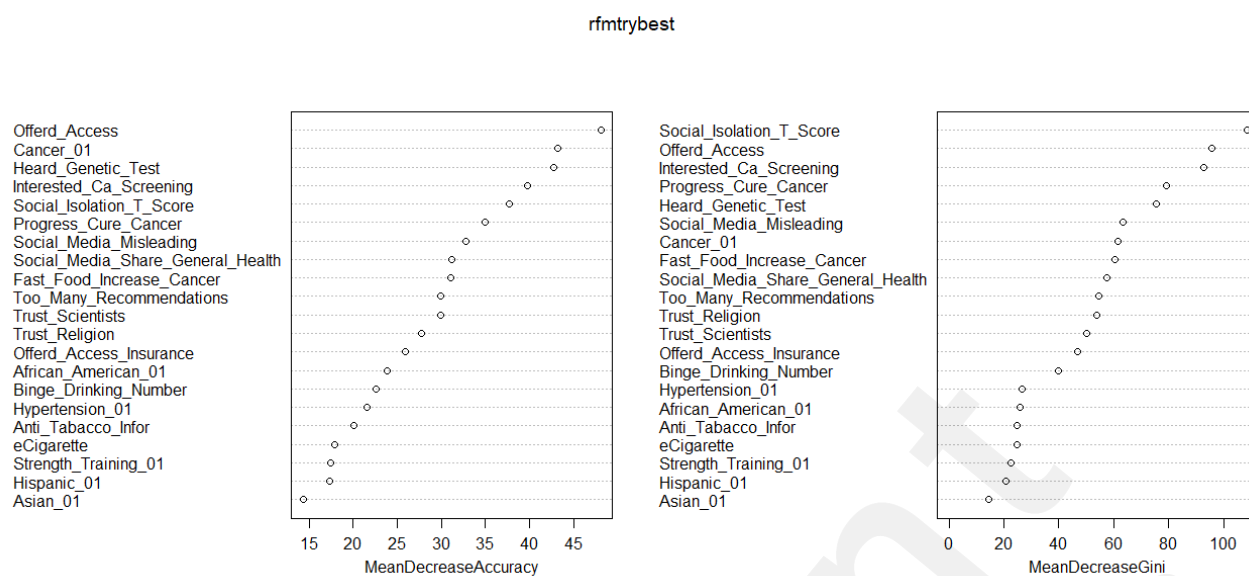


Figure 4. Plot of mean decrease accuracy (left panel) and mean decrease gini (right panel) values using random forest algorithm and 21 variables with highest loading from 21 factors

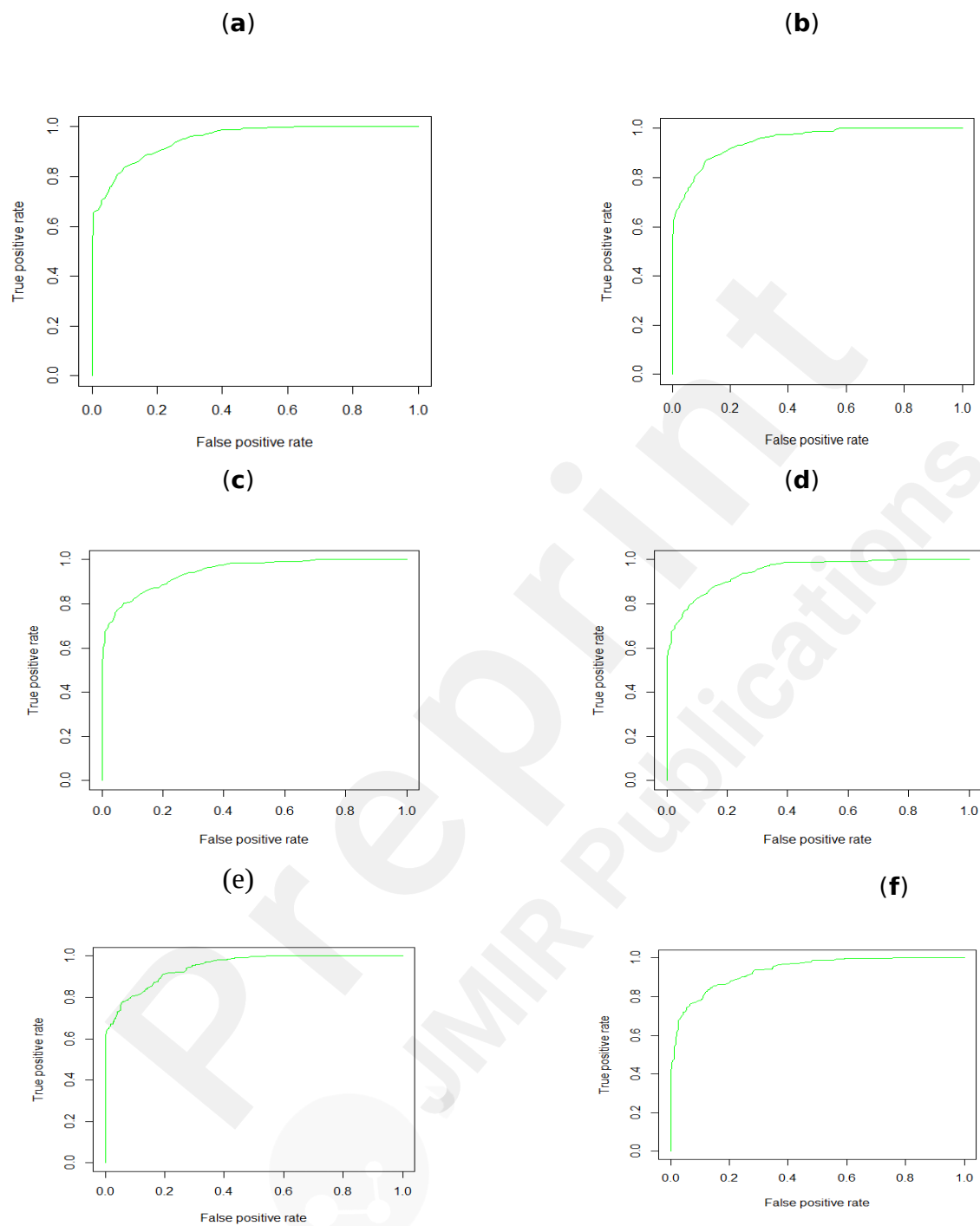


Figure 5. AUC curves in the validation data using random forest. (a) V73, (b) LASSO, (c) Boruta, (d) LASSO and Boruta, (e) 21 factors based on PCA score, (f) 21 variables with highest loading from 21 factors using PCA

Supplementary Files