Performance of the Classical Model in Feature Selection Across Varying Database Sizes of Healthcare Data

Kannan Thiruvengadam[1], Dadakhalandar Doddamani[2] and Rajendran Krishnan[1,*]

[1]ICMR – National Institute for Research in Tuberculosis, Chennai, India

[2]ICMR – Regional Medical Research Centre, Port Blair, Port Blair, India

Abstract: Machine learning is increasingly being applied to medical research, particularly in selecting predictive modelling variables. By identifying relevant variables, researchers can improve model accuracy and reliability, leading to better clinical decisions and reduced overfitting. Efficient utilization of resources and the validity of medical research findings depend on selecting the right variables. However, few studies compare the performance of classical and modern methods for selecting characteristics in health datasets, highlighting the need for a critical evaluation to choose the most suitable approach. We analysed the performance of six different variable selection methods, which includes stepwise, forward and backward selection using p-value and AIC, LASSO, and Elastic Net. Health-related surveillance data on behaviors, health status, and medical service usage were used across ten databases, with sizes ranging from 10% to 100%, maintaining consistent outcome proportions. Varying database sizes were utilized to assess their impact on prediction models, as they can significantly influence accuracy, overfitting, generalizability, statistical power, parameter estimation reliability, computational complexity, and variable selection. The stepwise and backward AIC model showed the highest accuracy with an Area under the ROC Curve (AUC) of 0.889. Despite its sparsity, the Lasso and Elastic Net model also performed well. The study also found that binary variables were considered more crucial by the Lasso and Elastic Net model. Importantly, the significance of variables remained consistent across different database sizes. The study shows that no major variations in results between the fitness metric of the model and the number of variables in stepwise and backward p-value models, irrespective of the database's size. LASSO and Elastic Net models surpassed other models throughout various database sizes, and with fewer variables.

Keywords: Models, regression analysis, machine learning, data interpretation, variable selection.

INTRODUCTION

The healthcare sector is undergoing rapid evolution, largely fuelled by the proliferation of electronic database applications and an increasing volume of data generated through health research. This transformation has led to the creation of large-scale databases, which provide significant resources for advanced data analytics [1]. By leveraging these databases, sophisticated analytical techniques that incorporate a variety of predictive algorithms and statistical software enable the development of advanced predictive models. A comprehensive understanding of the nuances associated with each predictive method, coupled with clearly defined model selection strategies, is crucial for conducting effective analyses [2].

For predictive models to achieve robustness, both internal and external validation are essential. These validation processes ensure that the most relevant evidence is employed, particularly in the context of data-driven decision-making [2]. To determine the most appropriate methodologies for prediction exercises, it is imperative to benchmark the performance of commonly used classical predictive techniques against healthcare datasets, particularly given their resource-intensive nature [3].

Prediction plays a fundamental role in the analysis of large datasets, with significant advancements made in the development of robust disease prediction models that utilize large-scale electronic databases and a combination of classical and naive methods [4]. A critical aspect of this model development process is variable selection, which significantly impacts model precision by refining inputs and eliminating unnecessary variables. However, reliance on literature-based or expert-driven variable selection can introduce bias, potentially obstructing the identification of novel insights within existing findings [5].

The variable selection process is vital in developing effective predictive models, focusing on reducing the number of variables by eliminating those with weak or non-informative relationships to the outcome. Failing to implement variable selection can lead to several issues, including overfitting, which undermines the model's generalizability, as well as increased computational complexity, resulting in longer training times and greater resource consumption. Additionally, models that incorporate irrelevant features may exhibit decreased accuracy, suffer from the curse of dimensionality, face challenges in interpretation, and demonstrate instability and bias [5-7].

*Address correspondence to this author at the Department of Epidemiology, ICMR-National Institute for Research in Tuberculosis, No. 1, Mayor Sathyamoorthy Road, Chennai – 600 031, India; Tel: +91 94459 22851;
E-mail: tkannan1985@gmail.com

To enhance model performance, interpretability, and efficiency, careful variable selection is essential. This process can be guided by the principle of parsimony, which advocates for simpler models over more complex ones, positing that explanations with fewer variables are generally preferable [8]. The prevailing view is that a model with non-significant variables removed is more parsimonious than a full model containing a greater number of predictors, thus justifying the preference for streamlined models in statistical practice [8].

Furthermore, to ensure accuracy and enable comprehensive comparisons of a single algorithm's performance, it is crucial to consider various study conditions. Neglecting these factors may introduce biases and yield misleading results [9]. The size of the database and the ratio of events per variable also play critical roles in the performance of models across various classical and naive approaches during the development of accurate prediction models [10].

This study aims to provide an in-depth assessment of the relative performance of various algorithmic regression techniques for outcome prediction. By offering a critical understanding of the relative performance of these methodologies, this research will assist researchers in selecting appropriate algorithms, ultimately contributing to improve predictive modelling.

METHODS

We evaluated the performance of various modelling techniques, including stepwise selection with a p-value stopping rule, backward selection with a p-value stopping rule, stepwise selection based on the lowest Akaike Information Criterion (AIC), backward selection based on the lowest AIC, the Least Absolute Shrinkage and Selection Operator (LASSO), and the Elastic Net approach.

Stepwise and backward selection using p-values is a classical variable selection that has been extensively used by the researchers. In this method, all the variables will be used, and subsequently, the ones with the least significance will be eliminated. The stopping rule will be less than a 5% significance level while using the p-value [11] as criteria. The stepwise and backward selection using AIC is similar to the p-value methods except for the stopping rule, which was based on achieving the lowest information criterion using AIC [12].

Lasso regression combines shrinkage with variable selection by applying penalties to regression coefficients. It introduces the L1 norm for the problem of least squares, reduces the coefficients to zero, and is therefore eliminated from the model. This difference in model metrics may seem minor, but it has a major impact on the model validity [13]. To find the optimal constraint parameter in our implementation, we performed 100 sequential searches over a parameter grid of 0.02 increments and calculated the area under the receiver operating characteristic curve (AUC) using ten-fold cross-validation. Elastic Net is an extension of Lasso where L1 and L2 penalties are imposed. The characteristics of the L2 standard promote group effects, allowing highly correlated variables to be retained in the model or eliminated together in a structured manner similar to Lasso. An elastic network can better manage situations in which the number of predictors exceeds the number of cases [14]. In our implementation, the optimal parameters were found by performing 100 random searches over a parameter grid and calculating the AUC using a two-dimensional ten-fold cross-validation. The models of these methods achieving the highest AUC was employed to identify selected variables and assess method performance on the validation set.

Datasets

The Behavioral Risk Factor Surveillance System (BRFSS) is a thorough telephone survey designed to gather self-reported information from adult residents aged 18 and older regarding their health-related behaviors, overall health status, and use of medical services. For further information on the survey's methodology and data, please refer to the detailed explanation provided elsewhere [15].

The 2022 BRFSS data was utilized for predictive modelling of general health status. A total of 34 variables were selected for analysis, encompassing factors such as residential area, age, marital status, gender, body mass index, education, income, occupation, physical activity, days without good physical or mental health, sleep duration, depression status, access to healthcare, and various comorbidities including stroke, cancer, COPD, kidney problems, diabetes, asthma, arthritis, heart disease, COVID, and disabilities like deafness, blindness, concentration difficulties, walking difficulties, smoking habits, and alcohol consumption. Prior to analysis, all missing values were removed, and every dataset was divided into a 75/25 split for training and validating the model, ensuring that the outcome proportion was the same in both sets [16].

For analysis, we created nine additional databases with data ranging from 10% to 90%, maintaining the same proportion of 16.8% Poor Health vs 83.2% Good Health. Database sizes varied from 25083 to 225746, with the total dataset consisting of 250829 records.

Method Evaluation

The evaluation of the methods was conducted based on specific criteria such as parsimony, performance changes in relation to the reference model, and variable importance during selection. Previous research has indicated that parsimony plays a crucial role in enhancing prediction accuracy [8,17], and it was used to assess the trade-off between sparsity and prediction accuracy for each method. Performance of the validation was performed by assessing the AUC and the number of variables chosen by the approach was used to quantify sparsity.

Reference models were established for each method using all available variables in the dataset, and comparison models were created using the selected variables based on the parsimony exercise. The change in performance from the reference model was evaluated for each method using the method-specific default settings, in line with previous literature comparing modelling methods. The procedure for assessing the importance and selection of variables involved ranking each variable according to its significance. Each variable was then assigned a category, ranging from first to fourth, based on its importance quartile, the variables with the highest importance being represented by the fourth quartile. With zero importance denoting the non-selection of that particular variable, this approach was chosen to allow for comparison between methods. In all evaluations conducted, 75% of the data was used for deriving methods, while the remaining 25% was reserved for validating the metrics. This approach ensured a comprehensive assessment of the methods' performance using a representative subset of the original dataset.

Analysis

STATA version 16.0 (StatCorp, College Station, TX) and R version 4.2.2 (R Foundation for Statistical Computing, Vienna, Austria) were used to analyze the data. For metric estimation purposes, the R packages glmnet, VSURF, gbm, and caret were used [18].

RESULTS

Out of 250829 records, 208687 (83.2%) had better health outcomes and 42142 (16.8%) had poor health outcomes. The preliminary analysis in Table **1** demonstrates that all listed features were significantly associated with the health outcomes of the data.

Further analysis was performed using six different models, and their performance was shown in Figure **1** and Supplementary Figure **1**. These figures show the area under the curve against the number of features selected. Overall, the most accurate model in the dataset was stepwise and backward AIC with 28 variables and an AUC of 0.889. Lasso and Elastic Net had 26 variables with an AUC of 0.888, making it the sparsest model. The least performing models were stepwise and backward p-value with 30 variables and an AUC of 0.886 (Figure **1**).

The performance of these six models was similar in the varying database size as it was in the whole dataset (Supplementary Figure **1**).

The AUC improved and stabilized as the volume of the database for the stepwise and backward AIC models, whereas both the AUC and the number of variables improved and stabilized as the volume of the dataset increased (Figure **2**).

The reference models for the complete dataset had an AUC of 0.886 in the validation set for the stepwise and backward AIC & p-value models, whereas 0.882 for the Lasso and Elastic Net models. In evaluating model performance, Elastic Net, LASSO, stepwise, and backward AIC models demonstrated superior results. This assessment was based on the performance changes measure, calculated by subtracting the AUC of the reference model from the AUC of the tested models. The comparison also took into account the number of variables selected for different database sizes (Figures **2** and Supplementary Figure **2**).

The performance of the Elastic Net and Lasso in the complete dataset showed a significant improvement when compared to their reference model, whereas there was no significant change in the performance in the AIC models, and the p-value models showed some loss of performance. Lasso and Elastic Net Model sparsity had a negligible impact on this relationship, implying that approaches with a smaller number of variables performed better than other approaches. This similar trend existed throughout the varying size of the database.

Table 1: The Data Profile by the Participants Health Status used for the Comparison of the Different Methods

Variable	Overall, N = 250,829	Good or Better Health, N = 208,687	Fair or Poor Health, N = 42,142	p-value[1]
Gender, n (%)				<0.001
Female	128,214 (51.1)	106,146 (50.9)	22,068 (52.4)	
Male	122,615 (48.9)	102,541 (49.1)	20,074 (47.6)	
Age, Median (IQR)	58 (42 – 70)	57 (41 – 69)	62 (50 – 72)	<0.001
Residential Area, n (%)				<0.001
Rural	33,290 (13.3)	26,721 (12.8)	6,569 (15.6)	
Urban	217,539 (86.7)	181,966 (87.2)	35,573 (84.4)	
Marital Status, n (%)				0.040*
Never Married	51,081 (20.4)	42,344 (20.3)	8,737 (20.7)	
Ever Married	199,748 (79.6)	166,343 (79.7)	33,405 (79.3)	
Body Mass Index (kg/m^2), n (%)				<0.001
Under-Weight (BMI<18.5)	3,532 (1.4)	2,666 (1.3)	866 (2.1)	
Normal (18.5≥BMI<25.0)	70,207 (28.0)	61,554 (29.5)	8,653 (20.5)	
Over-Weight (25.0≥BMI<30.0)	89,331 (35.6)	77,366 (37.1)	11,965 (28.4)	
Obese (BMI≥20.0)	87,759 (35.0)	67,101 (32.2)	20,658 (49.0)	
Education, n (%)				<0.001
Did not graduate High School	10,629 (4.2)	6,237 (3.0)	4,392 (10.4)	
Graduated High School	56,591 (22.6)	43,563 (20.9)	13,028 (30.9)	
Attended College or Technical School	69,550 (27.7)	56,272 (27.0)	13,278 (31.5)	
Graduated from College or Technical School	114,059 (45.5)	102,615 (49.2)	11,444 (27.2)	
Income, n (%)				<0.001
<$15K	12,676 (5.1)	6,852 (3.3)	5,824 (13.8)	
$15K to <$25K	22,485 (9.0)	14,504 (7.0)	7,981 (18.9)	
$25K to <$35K	28,612 (11.4)	21,264 (10.2)	7,348 (17.4)	
$35K to <$50K	33,168 (13.2)	26,797 (12.8)	6,371 (15.1)	
$50K to <$100K	80,418 (32.1)	70,555 (33.8)	9,863 (23.4)	
$100K to <$200K	55,561 (22.2)	51,630 (24.7)	3,931 (9.3)	
≥$200K	17,909 (7.1)	17,085 (8.2)	824 (2.0)	
Employed, n (%)				<0.001
Not Employed	115,382 (46.0)	86,338 (41.4)	29,044 (68.9)	
Employed	135,447 (54.0)	122,349 (58.6)	13,098 (31.1)	
Student, n (%)				<0.001
Non-Student	246,251 (98.2)	204,556 (98.0)	41,695 (98.9)	
Student	4,578 (1.8)	4,131 (2.0)	447 (1.1)	
Retired, n (%)				<0.001
Non-Retired	171,341 (68.3)	145,155 (69.6)	26,186 (62.1)	
Retired	79,488 (31.7)	63,532 (30.4)	15,956 (37.9)	
Sleep Time in Hours	7 (6 – 8)	7 (6 – 8)	7 (6 – 8)	<0.001
No of days with poor physical health	0 (0 – 3)	0 (0 – 2)	10 (0 – 30)	<0.001
No of days with poor mental health	0 (0 – 5)	0 (0 – 3)	2 (0 – 15)	<0.001

(Table 1). Continued.

Variable	Overall, N = 250,829	Good or Better Health, N = 208,687	Fair or Poor Health, N = 42,142	p-value[1]
Having health insurance, n (%)				<0.001
Do not have some form of health insurance	11,063 (4.4)	8,752 (4.2)	2,311 (5.5)	
Have some form of insurance	239,766 (95.6)	199,935 (95.8)	39,831 (94.5)	
Leisure Time Physical Activities, n (%)				<0.001
No physical activity or exercise in last 30 days	55,750 (22.2)	36,553 (17.5)	19,197 (45.6)	
Had physical activity or exercise	195,079 (77.8)	172,134 (82.5)	22,945 (54.4)	
Depression, n (%)				<0.001
Yes	54,260 (21.6)	38,171 (18.3)	16,089 (38.2)	
No	196,569 (78.4)	170,516 (81.7)	26,053 (61.8)	
Affordable to seek Doctor, n (%)				<0.001
Affordable to Consult Doctor	231,538 (92.3)	196,057 (93.9)	35,481 (84.2)	
Could not Afford, Doctor	19,291 (7.7)	12,630 (6.1)	6,661 (15.8)	
Stroke, n (%)				<0.001
Yes	10,611 (4.2)	5,902 (2.8)	4,709 (11.2)	
No	240,218 (95.8)	202,785 (97.2)	37,433 (88.8)	
Cancer, n (%)				<0.001
Yes	29,895 (11.9)	21,743 (10.4)	8,152 (19.3)	
No	220,934 (88.1)	186,944 (89.6)	33,990 (80.7)	
COPD, n (%)				<0.001
Yes	20,082 (8.0)	10,457 (5.0)	9,625 (22.8)	
No	230,747 (92.0)	198,230 (95.0)	32,517 (77.2)	
Kidney Problem, n (%)				<0.001
Yes	11,730 (4.7)	6,412 (3.1)	5,318 (12.6)	
No	239,099 (95.3)	202,275 (96.9)	36,824 (87.4)	
Diabetes, n (%)				<0.001
Non-Diabetes	210,207 (83.8)	182,760 (87.6)	27,447 (65.1)	
Pre-Diabetes	5,665 (2.3)	4,182 (2.0)	1,483 (3.5)	
Diabetes	34,957 (13.9)	21,745 (10.4)	13,212 (31.4)	
Asthma, n (%)				<0.001
No Asthma	213,657 (85.2)	181,664 (87.1)	31,993 (75.9)	
Past Asthma	10,939 (4.4)	8,939 (4.3)	2,000 (4.7)	
Current Asthma	26,233 (10.5)	18,084 (8.7)	8,149 (19.3)	
Arthritis, n (%)				<0.001
No Arthritis	161,667 (64.5)	144,163 (69.1)	17,504 (41.5)	
Arthritis	89,162 (35.5)	64,524 (30.9)	24,638 (58.5)	
Heart Disease, n (%)				<0.001
No Heart Disease	227,457 (90.7)	195,253 (93.6)	32,204 (76.4)	
Heart Disease	23,372 (9.3)	13,434 (6.4)	9,938 (23.6)	
COVID, n (%)				<0.001
Yes	80,692 (32.2)	67,658 (32.4)	13,034 (30.9)	
No	170,137 (67.8)	141,029 (67.6)	29,108 (69.1)	

(Table 1). Continued.

Variable	Overall, N = 250,829	Good or Better Health, N = 208,687	Fair or Poor Health, N = 42,142	p-value[1]
Deaf, n (%)				<0.001
Yes	22,506 (9.0)	15,641 (7.5)	6,865 (16.3)	
No	228,323 (91.0)	193,046 (92.5)	35,277 (83.7)	
Blind, n (%)				<0.001
Yes	12,331 (4.9)	6,897 (3.3)	5,434 (12.9)	
No	238,498 (95.1)	201,790 (96.7)	36,708 (87.1)	
Difficulty in Concentration, n (%)				<0.001
Yes	27,755 (11.1)	15,885 (7.6)	11,870 (28.2)	
No	223,074 (88.9)	192,802 (92.4)	30,272 (71.8)	
Difficulty in Walking, n (%)				<0.001
Yes	38,098 (15.2)	17,883 (8.6)	20,215 (48.0)	
No	212,731 (84.8)	190,804 (91.4)	21,927 (52.0)	
Smoking Status, n (%)				<0.001
Non-Smoker	226,627 (90.4)	192,430 (92.2)	34,197 (81.1)	
Current Smoker	14,857 (5.9)	9,797 (4.7)	5,060 (12.0)	
Past Smoker	9,345 (3.7)	6,460 (3.1)	2,885 (6.8)	
Heavy Alcohol, n (%)				<0.001
Non-Alcohol	232,947 (92.9)	193,224 (92.6)	39,723 (94.3)	
Alcohol	17,882 (7.1)	15,463 (7.4)	2,419 (5.7)	

[1] Mann Whitney test for the continuous observations; Fisher Exact and or Chi-square test was used for categorical observations.

The table presents a detailed data profile based on the prediction class, which represents the participants' health status in this study. It includes the frequency and column percentage for each variable, as well as the statistical association between the variables and health status. This information provides valuable insights into the relationship between the predictors and the outcome, allowing for a deeper understanding of the factors influencing health status.

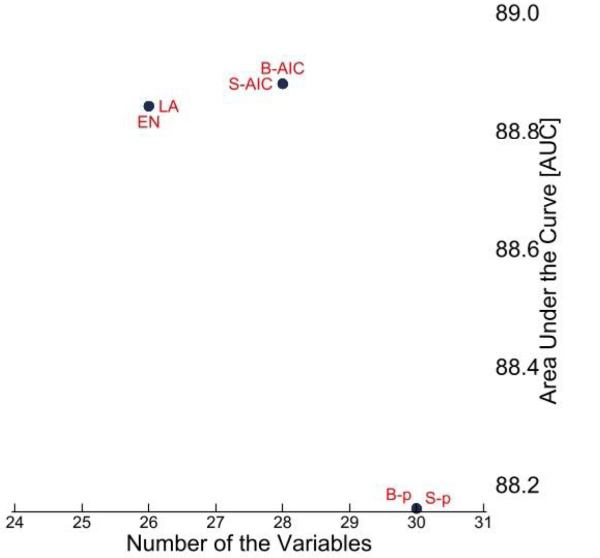

Figure 1: Parsimony measures, the area under the curve is the discrimination performance on the validation set.

A higher AUC value indicates better discrimination performance, meaning the model is able to effectively distinguish between the classes. On the other hand, a model with a higher AUC value and fewer variables is considered more robust as it is able to achieve good performance with less complexity.

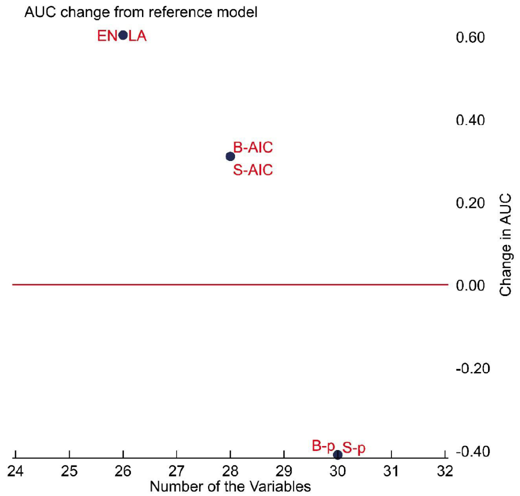

Figure 2: An evaluation of the discrimination ability of a model showing the variance in AUC between the full variable model and the reduced variable model.

The discrimination performance of a model was evaluated by comparing the difference in Area Under the Curve (AUC) between a model with all variables and a model with fewer variables from variable selection. A change greater than zero indicates improved performance, while a change less than zero indicates decreased performance. A model with fewer variables is considered more robust as it achieves good performance with less complexity.

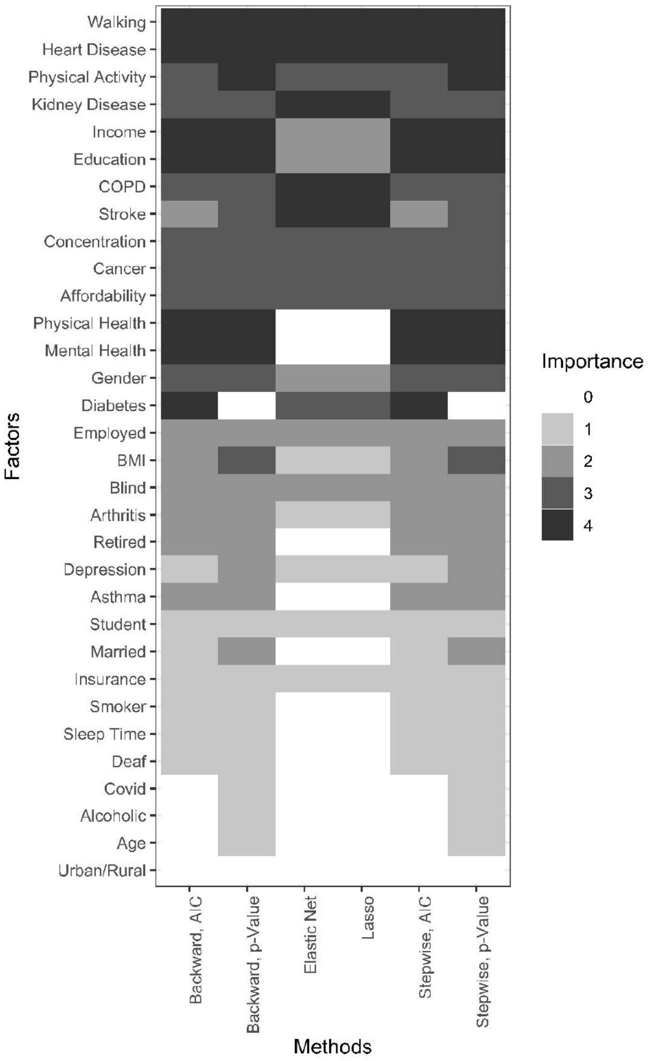

Figure 3: Heat map illustrating the predictive ability importance of variables based on their selection method.

The heat map provides a visual representation of the relationship between variables and feature selection methods, highlighting the importance of each method for each specific variable. By standardizing the importance metrics using quartiles, the heat map allows for easy comparison of the performance between variables and methods. This analysis aids in understanding the predictive ability and importance of variables in relation to the selection method.

The selected variables were ranked from the highest (4th quartile) to lowest (1st quartile) based on their importance, and the not included variables as zero for the whole dataset (Figure **3**), and the same has been shown for all the size varying databases in the Supplementary Figure **3**.

The Lasso and Elastic Net Model were found to be a sparsest models and they considered binary variables as higher importance in comparison to other models, as they did not rank continuous variables like Physical and Mental health as important, whereas all other four models highlighted it as a key variable. Walking ability and heart disease have consistently emerged as the most significant factors, consistently ranking in the highest quartile across different methods.

The importance of the variables was similar and consistent throughout the varying size of the database (Figure **4**). This indicates the strong and stable relationship between these variables regardless of the size of the dataset.

DISCUSSION

Varying dataset sizes were utilised to assess the performance of the different variable selection methods. There are no significant performance differences between model fitness metric and number of variables in stepwise and backward p-value models, regardless of database size. The performance of stepwise and backward AIC models improves with larger database sizes, along with the number of variables. Elastic Net and Lasso models outperform other models across different database sizes, with Lasso and Elastic Net performing better with fewer variables. However, these stepwise and backward techniques were overshadowed by few selection methods, such as recursive feature elimination with cross- validation [19,20] and regularized tree ensembles [21], those were addressing the curse of dimensionality and avoid overfitting. Those were also aimed to classify a parsimonious set of predictors and are confirmed with out-of-sample data for successful feature selection.

In stepwise regression models, the use of theoretical arguments or expert opinion to select initial predictors has been shown to be valuable. It is more successful to start with at least five to ten true variables and five to ten nuisance variables rather than starting with hundreds of nuisance variables. A Bayesian method is recommended, which integrates data with a prior distribution for the model's parameters to account for uncertainty in the relevance of potential predictors. As the amount of data increases, Bayesian posterior distributions increasingly prioritize error minimization, converging towards least squares estimates. This reduces uncertainty and improves precision, but may simultaneously mask the true variability inherent in the underlying data [22,23].

For explanatory variable coefficients, the LASSO, Elastic Net, and Ridge regression methods all implicitly assume independent, zero-mean, identical-variance

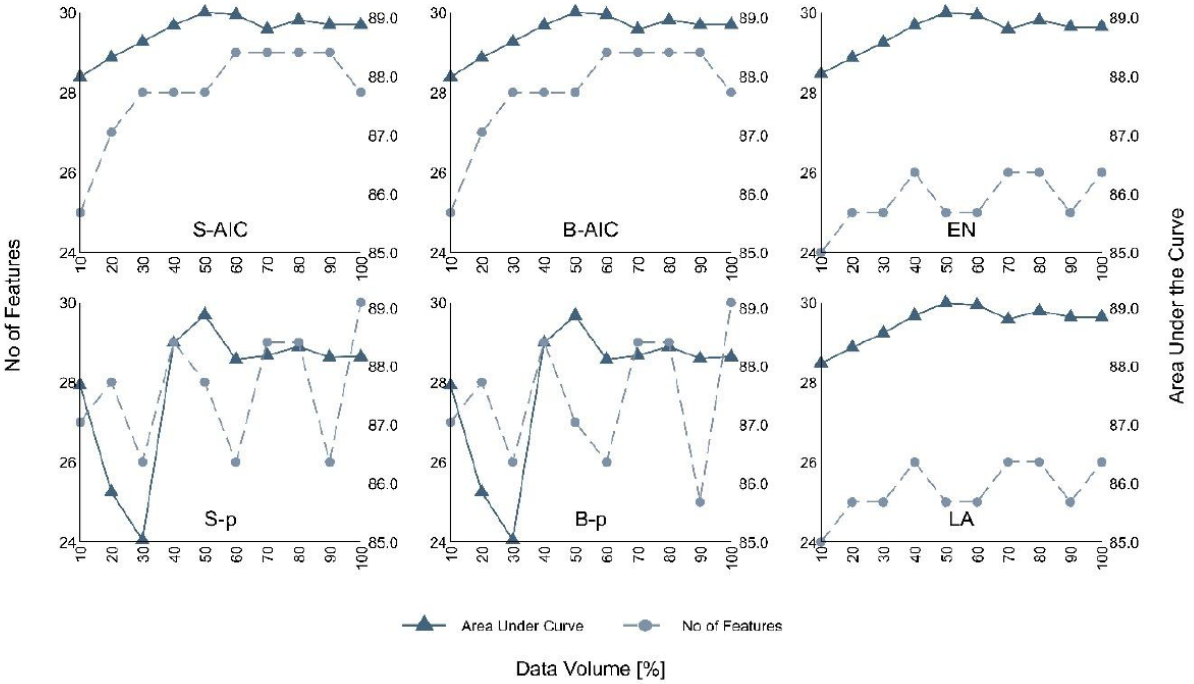

Figure 4: The relationship between the area under the curve and the number of features extracted depends on the size of the database.

The figure illustrates the correlation between the area under the curve and the number of features extracted by different methods, highlighting the impact of database size on this relationship. Despite the fluctuations in database size, the figure demonstrates a level of similarity and consistency in results for a specific method. This suggests that the method is robust and reliable across varying database sizes.

prior distributions[24]. However, it is implausible that predictor coefficients determined based on expert judgment would have previous means of zero. Explicit priors may be more appealing than implicit priors. The performance metrics of a model are influenced by the feature selection procedure and the classifier used. Certain combinations may excel with small sample sizes, while others may perform better with larger samples and show lower variance. It is important to carefully consider the selection of features and classifier to optimize model performance [25]. It is crucial to acknowledge that our investigation was limited to a select number of combinations, indicating that the results may not be directly transferable to variables obtained from data with non-Gaussian class distributions in smaller databases.

CONCLUSIONS

In summary, the performance of the model fitness metric and the number of variables extracted in the stepwise and backward p-value models do not show significant differences across different database sizes. The AUC performance of the stepwise and backward AIC models improves with larger databases and more variables, demonstrating some consistency. The Lasso and Elastic Net models consistently outperform other models across varying database sizes. While the AUC performance of the stepwise and backward AIC models, as well as the Lasso and Elastic Net models, show similar trends with database size, the Lasso and Elastic Net models perform better with fewer variables. Further research is needed to assess the generalizability of these findings to different clinical scenarios and datasets.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

AVAILABILITY OF DATA AND MATERIAL

The original data for the study were included in the article as an open access dataset, which is available freely in the mentioned references. The codes used for the analysis will be shared upon reasonable request directed to the corresponding author.

COMPETING INTEREST

The authors declare that they have no competing interests.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

AUTHORS CONTRIBUTIONS

The study concept, design, data accumulation, analysis, interpretation, and critical evaluation were all contributions made by KT and RK. Manuscript writing and corrections were also completed by KT, DD and RK. The final article received approval from all authors, who read and approved the final manuscript.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the ICMR – National Institute for Research in Tuberculosis for their exceptional support and granting us permission to conduct our research activities at their facilities.

LIST OF ABBREVIATIONS

AIC	=	Akaike Information Criterion
AUC	=	Area under the ROC Curve
BRFSS	=	Behavioral Risk Factor Surveillance System
COPD	=	Chronic obstructive pulmonary disease
Lasso	=	Least Absolute Shrinkage and Selection Operator
p-value	=	Probability value, is a number that describes the likelihood of a result occurring under the assumption of no effect or no difference
ROC	=	Receiver Operating Characteristic

SUPPLEMENTARY FIGURES

The supplementary figures can be downloaded from the journal website along with the article.

REFERENCES

[1] Evans RS. Electronic Health Records: Then, Now, and in the Future. Yearb Med Inform 2016; 25. https://doi.org/10.15265/IYS-2016-s006

[2] Brnabic A, Hess LM. Systematic literature review of machine learning methods used in the analysis of real-world data for patient-provider decision making. BMC Medical Informatics and Decision Making 2021; 21(1): 1-19. https://doi.org/10.1186/s12911-021-01403-2

[3] van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. BMC Medical Research Methodology 2014; 14(1): 137-137. https://doi.org/10.1186/1471-2288-14-137

[4] Hossain E, Hossain E, Khan A, Moni MA, Uddin S. Use of Electronic Health Data for Disease Prediction: A Comprehensive Literature Review. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2019; 18(2): 745-58. https://doi.org/10.1109/TCBB.2019.2937862

[5] Bagherzadeh-Khiabani F, Ramezankhani A, Azizi F, Hadaegh F, Steyerberg EW, Khalili D. A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. Journal of Clinical Epidemiology 2016; 71: 76-85. https://doi.org/10.1016/j.jclinepi.2015.10.002

[6] Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Family Medicine and Community Health 2020; 8(1). https://doi.org/10.1136/fmch-2019-000262

[7] He L, He Lingjun, Levine RA, Fan J, Beemer J, Stronach J. Random Forest as a Predictive Analytics Alternative to Regression in Institutional Research. Practical Assessment, Research and Evaluation 2018; 23(1): 1.

[8] Steyerberg EW. Clinical Prediction Models 2009. https://doi.org/10.1007/978-0-387-77244-8

[9] Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 2006; 7(1): 1-30.

[10] Rodriguez D, Catal C, Cagatay Catal, Diri B. Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem. Information Sciences 2009; 179(8): 1040-58. https://doi.org/10.1016/j.ins.2008.12.001

[11] Murtaugh PA. In defense of P values. Ecology 2014; 95(3): 611-7. https://doi.org/10.1890/13-0590.1

[12] Portet S. A primer on model selection using the Akaike Information Criterion. Infectious Disease Modelling 2020; 5: 111-28. https://doi.org/10.1016/j.idm.2019.12.010

[13] Tibshirani R. Regression shrinkage and selection via the lasso: a retrospective. Journal of The Royal Statistical Society Series B-statistical Methodology 2011; 73(3): 273-82. https://doi.org/10.1111/j.1467-9868.2011.00771.x

[14] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of The Royal Statistical Society Series B-statistical Methodology 2005; 67(2): 301-20. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[15] Control C for D, Prevention, others. Behavioral risk factor surveillance system survey questionnaire. Atlanta, Georgia: US Department of Health and Human Services, Centers for Disease Control and Prevention 2022; 22-3.

[16] Control C for D, Prevention, others. Behavioral risk factor surveillance system survey data. http://appsnccdcdc gov/brfss/listasp?cat=OH&yr-2008&qkey=6610&state=All 2022.

[17] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference and prediction. Math Intell 2005.

[18] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria 2022. [Internet]. Available from: https://www.R-project.org/

[19] Mukherjee T, Mukherjee T, Duckett M, Kumar P, Paquet J, Paquet JD, *et al.* RSSI-Based Supervised Learning for Uncooperative Direction-Finding. ECML/PKDD 2017; 216-27. https://doi.org/10.1007/978-3-319-71273-4_18

[20] Guyon I, Jason Weston, Weston J, Barnhill S, Vladimir Vapnik, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 2002; 46(1): 389-422.
https://doi.org/10.1023/A:1012487302797

[21] Deng H, George C. Runger, Runger GC. Feature Selection via Regularized Trees. arXiv: Learning 2012.

[22] Gelman A, John B. Carlin, John B. Carlin, Carlin JB, Hal S. Stern, Stern HS, *et al*. Bayesian data analysis, third edition 2013.
https://doi.org/10.1201/b16018

[23] Box GEP, George C. Tiao, Tiao GC, Tiao GC. Bayesian Inference in Statistical Analysis: Box/Bayesian 1992.
https://doi.org/10.1002/9781118033197

[24] Smith G, Frank Campbell, Campbell F. A Critique of Some Ridge Regression Methods. Journal of the American Statistical Association 1980; 75(369): 74-81.
https://doi.org/10.1080/01621459.1980.10477428

[25] Way TW, Berkman Sahiner, Sahiner B, Hadjiiski LM, Chan HP. Effect of finite sample size on feature selection and classification: a simulation study. Medical Physics 2010; 37(2): 907-20.
https://doi.org/10.1118/1.3284974

Received on 19-08-2024 Accepted on 16-09-2024 Published on 14-10-2024

https://doi.org/10.6000/1929-6029.2024.13.21

© 2024 Thiruvengadam *et al*.

This is an open-access article licensed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the work is properly cited.