Original Paper

A Risk Prediction Model for Physical Restraints Among Older Chinese Adults in Long-term Care Facilities: Machine Learning Study

Jun Wang[1*], MNS; Hongmei Chen[2*], MNS; Houwei Wang[3], BS; Weichu Liu[2], MNS; Daomei Peng[4], BSN; Qinghua Zhao[1], BSN; Mingzhao Xiao[5], MM

[1]Department of Nursing, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

[2]Department of Gynecology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

[3]College of Mathematics and Physics, Chongqing University of Science and Technology, Chongqing, China

[4]Aged Care Unit, The First Social Welfare Home of Chongqing, Chongqing, China

[5]Department of Urology, The First Affiliated Hospital of Chongqing Medical University, Chongqing, China

[*]these authors contributed equally

Corresponding Author:
Mingzhao Xiao, MM
Department of Urology
The First Affiliated Hospital of Chongqing Medical University
1 Youyi Road, Yuzhong District
Chongqing, 400016
China
Phone: 86 13608399433
Email: xmz.2004@163.com

Abstract

Background: Numerous studies have identified risk factors for physical restraint (PR) use in older adults in long-term care facilities. Nevertheless, there is a lack of predictive tools to identify high-risk individuals.

Objective: We aimed to develop machine learning (ML)–based models to predict the risk of PR in older adults.

Methods: This study conducted a cross-sectional secondary data analysis based on 1026 older adults from 6 long-term care facilities in Chongqing, China, from July 2019 to November 2019. The primary outcome was the use of PR (yes or no), identified by 2 collectors' direct observation. A total of 15 candidate predictors (older adults' demographic and clinical factors) that could be commonly and easily collected from clinical practice were used to build 9 independent ML models: Gaussian Naïve Bayesian (GNB), k-nearest neighbor (KNN), decision tree (DT), logistic regression (LR), support vector machine (SVM), random forest (RF), multilayer perceptron (MLP), extreme gradient boosting (XGBoost), and light gradient boosting machine (Lightgbm), as well as stacking ensemble ML. Performance was evaluated using accuracy, precision, recall, an *F* score, a comprehensive evaluation indicator (CEI) weighed by the above indicators, and the area under the receiver operating characteristic curve (AUC). A net benefit approach using the decision curve analysis (DCA) was performed to evaluate the clinical utility of the best model. Models were tested via 10-fold cross-validation. Feature importance was interpreted using Shapley Additive Explanations (SHAP).

Results: A total of 1026 older adults (mean 83.5, SD 7.6 years; n=586, 57.1% male older adults) and 265 restrained older adults were included in the study. All ML models performed well, with an AUC above 0.905 and an *F* score above 0.900. The 2 best independent models are RF (AUC 0.938, 95% CI 0.914-0.947) and SVM (AUC 0.949, 95% CI 0.911-0.953). The DCA demonstrated that the RF model displayed better clinical utility than other models. The stacking model combined with SVM, RF, and MLP performed best with AUC (0.950) and CEI (0.943) values, as well as the DCA curve indicated the best clinical utility. The SHAP plots demonstrated that the significant contributors to model performance were related to cognitive impairment, care dependency, mobility decline, physical agitation, and an indwelling tube.

Conclusions: The RF and stacking models had high performance and clinical utility. ML prediction models for predicting the probability of PR in older adults could offer clinical screening and decision support, which could help medical staff in the early identification and PR management of older adults.

XSL•FO
RenderX

Introduction

Physical restraint (PR) is not only an important indicator to measure the quality of medical care but is also a major public health issue that has aroused widespread concern worldwide. PR is defined as "any action or procedure that prevents a person's free body movement to a position of choice or normal access to his or her body by the use of any method, attached or adjacent to a person's body that he or she cannot control or remove easily" [1]. It is usually used to protect patients from falls, self-extubation, or injuries in intensive care units or psychiatric hospitals. Recent studies have shown that PR is more widely applied among older adults and results in worse outcomes. Older adults are 3 times more likely to endure PR than young people during hospitalization [2]. A systematic review and meta-analysis indicated that the pooled prevalence of PR among older adults in long-term care (LTC) facilities ranges from 22% (in North America) to 65% (in Australia) [3]. In China, the prevalence of PR among older adults in LTC was 62% in Taiwan [4], 52.7%-70.2% in Hong Kong [5], and 25.8% in mainland China (Chongqing) [6].

The World Health Organization has reported that restraining individuals could be considered maltreatment [7]. As older adults are vulnerable to various health-related problems, PR is used more frequently and lasts longer, resulting in more serious injuries. Previous studies have demonstrated that PR is inadequate for protection [8-10]. Conversely, PR is associated with negative consequences on physical (eg, pressure ulcers, fractures, and urinary and fecal incontinence), psychological (eg, cognitive decline, depression, anxiety, aggression, and fear), and social functions (eg, social isolation and loss of social worth), and even death [11-16]. Therefore, the early identification of high-risk individuals and early interventions are of great significance in preventing PR use, which can reduce the negative impact on health, society, and the economy.

Numerous studies have identified that the risk factors of PR use are associated with (1) individual-related factors (eg, age, cognitive impairment, mobility decline, care dependency, fall risk, etc) [16-18], (2) facility-related factors (eg, facility type, ownership, and staff levels) [6,17,19], and (3) caregiver-related factors (eg, knowledge, attitude, and intention) [20-22]. Nevertheless, the interaction between these predictors and their clinical value remains unclear. Moreover, predictors of PR use are mainly determined using traditional statistical methods, with the limitations of difficulty dealing with high-dimensional data, nonlinear variables, and heterogeneous distribution [23,24]. Importantly, rare tools, such as the model proposed in this paper, can comprehensively assess restraint risks and support decision-making for staff and families. The clinical prediction model provides a new horizon. This would allow early detection and increased surveillance of at-risk older adults and the development of early targeted interventions for preventing and reducing PR.

Recently, machine learning (ML) algorithms such as Naïve Bayes and random forest (RF) have been used in various fields for clinical practice such as diagnosis, occurrence, and prognosis [25-27]. Various robust ML prediction models have been developed for adverse events and complications prediction, such as cognitive impairment prediction [28], falls prediction [29], pressure injury prediction [30], and delirium prediction [31]. Current regular risk prediction models were developed using a generalized linear model, which depends on the implicit assumption that each risk factor is related in a linear fashion to outcomes. Although such a model is easy to code and fast to calculate, it may oversimplify the complex nonlinear interaction between variables. ML methods have been an alternative method to address current limitations. ML helps in handling information based on causal or statistical data, potentially revealing hidden dependencies between factors and diseases, and supporting clinical decisions [32]. To our knowledge, the development of a PR prediction model using ML has only been studied in psychiatric inpatients, limiting its predictive performance and application to broader scenarios [33,34]. Moreover, there are few studies on PR prediction among older adults in LTC facilities.

Given that our previous multicenter investigative study identified risk factors for PR, we have now proceeded to develop a prediction model for PR with information that can easily be provided by older adults in LTC facilities. The purpose of this study was to (1) develop and compare 9 independent ML models, (2) analyze the most important features of the 2 models with the best prediction performance, and (3) train and validate a more stable and generalized stacked model using the stacking ensemble learning algorithm.

Methods

Study Design and Participants

This study comprises a secondary analysis of multicenter cross-sectional data from 6 LTC facilities (ie, 1 aged-care center, 1 social welfare home, and 4 nursing homes) in Chongqing, China, from July 2019 to November 2019. Based on the inclusion criterion, all older adults who were present in the LTC facilities on the days of data collection were approached. The exclusion criteria were as follows: (1) older adults who have lived in LTC facilities for less than 2 weeks and (2) older adults whom we were not allowed to observe because of serious and special illnesses, brain death, or no voluntary movement ability. This study followed the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis reporting guidelines [35].

Ethics Approval

The study was approved by the Ethics Committee of the First Affiliated Hospital of Chongqing Medical University (approval number: 2019-104). Informed consent was obtained from all participants in the primary data collection, and the original informed consent allows the secondary analysis without additional consent. All research data are anonymous or deidentified to protect the privacy of participants. We did not provide any remuneration to each participant who provided complete and valid responses.

Outcome and Predictors

The primary outcome was PR use (yes or no), gathered by 2 collectors through 3 direct observations, which was reported to be the most reliable method for collecting data on PR use [36]. The 3 separate observations were conducted on a working day at times when older adults were either most likely to be active or at rest. A "yes" response was recorded if at least one PR was used during the 3 data collection periods. The definition of PR complied with an international consensus [1]. We excluded bed rails as a means of PR because bed rails are conventionally pulled up in these facilities when older adults are lying in bed, but they can be removed freely as needed.

Predictors were identified based on a literature review and the clinical knowledge of practitioners who worked in LTC facilities. Details of the measurement and data collection are available in a previously published article [6]. A total of 15 predictors related to individual factors were included in this study: (1) older adults' sociodemographic characteristics (ie, sex, age, and length of residence at LTC); and (2) clinical factors such as functional status on present living, including chronic diseases, mental diseases, consciousness, cognitive function, mobility, degree of care dependency, physical agitation, verbal agitation, depression symptoms, fecal and urinary conditions, fall risk, and indwelling tube. The definition of each variable was presented in Table S1 in Multimedia Appendix 1.

Statistical Analysis

Overview

The demographic characteristics of older adults were described using descriptive statistics, such as means, SD, numbers, and percentages. The chi-square test was performed to compare predictive variables between the PR and non-PR groups. These analyses were performed using IBM's SPSS, version 25.0. All statistical tests were 2-sided, and a P value less than .05 was regarded as statistically significant.

For the ML model development, we used a 2-step systematic framework comprising 9 widely applied independent ML methods and stacked ensemble-based ML models. Nine independent ML methods were used: logistic regression (LR), Gaussian Naïve Bayes (GNB), k-nearest neighbors (KNN), support vector machine (SVM), decision tree (DT), RF, extreme gradient boosting (XGBoost), light gradient boosting machine (Lightgbm), and multilayer perceptron (MLP). ML algorithms were performed in Python (version 3.7.4; Guido van Rossum). An overall flowchart of the analyses is shown in Figure 1.

This study was a secondary analysis of cross-sectional data that assessed the risk of PR without missing values or outliers. For data preparation, the data set was randomly divided into a training set (80%) and a test set (20%), and the samples were added with the same nonzero random seeds and stratified to ensure that the proportions of the cases in the training and test sets were equal and improved the stability of the model. The data set was split only once into a training set and a test set for each of the 9 independent ML models. The performance of the model was evaluated based on accuracy, precision, recall, F score, and area under the receiver operating characteristic curve (AUC; based on the test set), which were widely used in other studies [37-39]. The 95% CI of the AUC was calculated with bootstrapping, using 1000 iterations. Moreover, a decision curve analysis (DCA) was performed to evaluate the clinical utility of the 5 best performance prediction models by quantifying the net benefits.

Figure 1. Flowchart summary of our methodology. AHP: analytic hierarchy process; DCA: decision curve analysis; DT: decision tree; GNB: Gaussian Naïve Bayesian; KNN: k-nearest neighbor; Lightgbm: light gradient boosting machine; LR: logistic regression; MLP: multilayer perceptron; RF: random forest; ROC: receiver operating characteristic; SHAP: Shapley Additive Explanations; SVM: support vector machine; XGBoost: extreme gradient boosting.
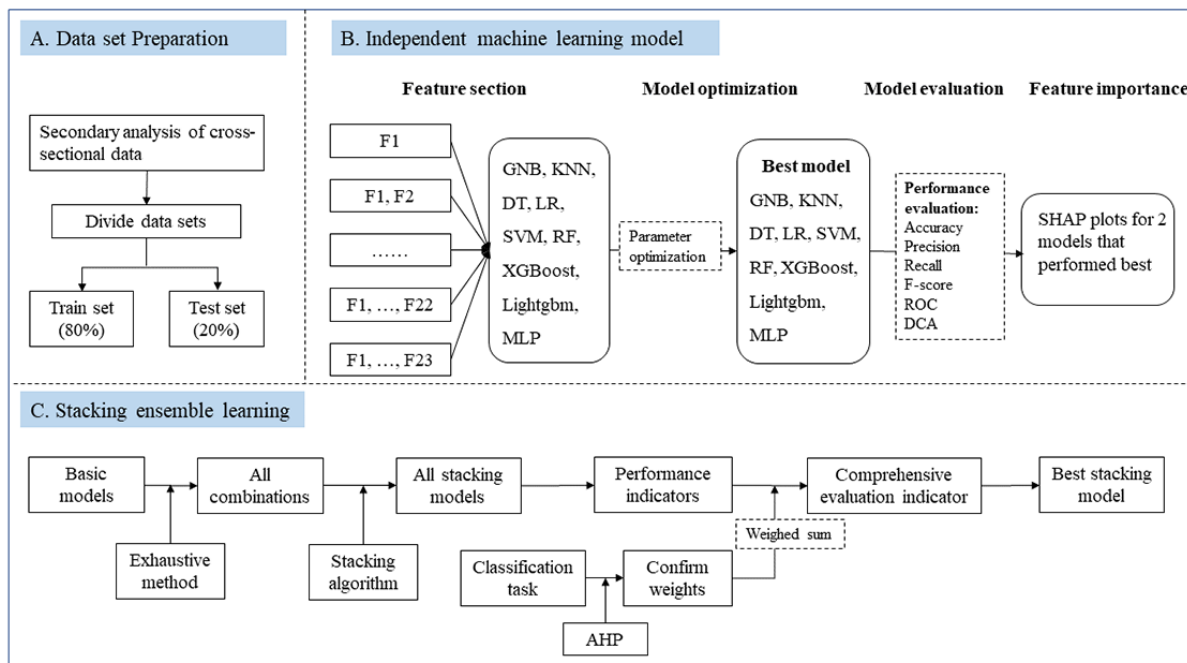
Independent ML Model Development

For each model, we ranked feature importance based on RF and subsequently entered 15 features in order of importance using the iterative screening method. The model performance for each feature combination was then recorded. The final feature selection was obtained based on the optimal number of features and accuracy. After feature selection, the hyperparameters of the model were optimized using the Python sklearn (David Cournapeau) cross-validated grid search function. Specifically, the parameters tuned in each model were searched and optimized individually. Iterative tuning was performed for each parameter within the parameter range, and a visualized learning curve was used to select the optimal parameter value to reach the local optimum. The above steps were repeated for each parameter tuned in the model, and the optimal range of the parameters in each model was determined. Finally, a grid search through 10-fold cross-validation was performed to ascertain the best parameter combination value of the model. The trained model was then validated on the test data set, with the output being evaluation indicators of the model's performance. We test the results via 10-fold cross-validation in the training set to avoid overfitting and assess the stability of the models. Furthermore, we used Shapley Additive Explanations (SHAP) to interpret and visualize the impact of predictors on PR risk based on the 2 models that performed best [40].

Stacking Ensemble Learning

We then used a stacking algorithm that has been shown to perform better than boosting and bagging ensemble classification algorithms [41,42]. The stacking-based algorithms contain cross-validation, which is used to select optimal basic model

parameters. In stacking-based models, we also test the results via 10-fold cross-validation in the training set to avoid overfitting and assess the stability of the models. The importance of the classifier performance evaluation indicators should also be different for different classification tasks. In this study, a comprehensive evaluation indicator (CEI) based on the purpose of the classification task was used to evaluate the stacking model's performance. It is defined as the weighted sum of the accuracy, precision, recall, and *F* score [43]. The weight of each indicator was determined using an analytic hierarchy process (AHP) [44]. In this study, we used the weight value calculated by Sun and Chen [43] with the values of accuracy (0.061), precision (0.293), recall (0.182), and *F* score (0.463). The weight value was calculated in a clinical setting to predict the occurrence of diseases, which was consistent with the requirements of our study prediction task. The AHP-stacking algorithm can be divided into 3 steps [43]: (1) list all possible basic model combinations using the exhaustive method based on the basic independent models, subsequently developing stacking models and output performance indicators; (2) determine the weights of the model performance indicators using AHP, based on classification tasks, and calculate the CEI of all stacking models; and (3) generate all stacking models and rank them in order of performance (ie, CEI).

Results

Participants' Demographics

A total of 1026 older adults in 6 LTC facilities were included, comprising 265 older adults in the restraint group and 761 older adults in the nonrestraint group. Figure S1 in Multimedia Appendix 1 presents a flowchart of participant selection.

Overall, the mean age was 83.47 (SD 7.62; range 60-102) years. A total of 586 (57.12%) participants were women. Table S2 in Multimedia Appendix 1 presents a univariate analysis of older adults with and without PR. Predictors comprising age, length of residence at LTC, number of chronic diseases, consciousness, cognitive function, mobility, care dependency, physical agitation, verbal agitation, depression symptoms, fecal and urinary conditions, fall risk, and indwelling tube showed significant differences between the PR and non-PR groups.

Model Evaluation and Performance

The feature selection results showed that GNB performed best with 9 features, while the other models incorporated all features. Table 1 presents the predictive performances of the 9 models. Among them, the RF model performed with the highest accuracy (0.922), followed by the SVM (0.903). DT and GNB ranked the lowest (0.859). The RF model performed well with respect to precision (0.953), followed by GNB (0.949), MLP (0.939), SVM (0.929), and others above 0.900. The SVM and RF models showed the greatest sensitivity, with a value of 0.941, and the GNB model showed the lowest sensitivity (0.856). Considering that precision and sensitivity are often contradictory, we calculated the F score, an evaluation indicator that weighed precision and sensitivity. The top 3 F score models were RF (0.947), SVM (0.935), and LR (0.925). The AUC illustrated that SVM and RF had the best predictive performance, with AUC values of 0.949 (95% CI 0.911-0.953) and 0.938 (95% CI 0.914-0.947), respectively. The other models are 0.900 above. The details are shown in Figure 2. Overall, the best-performing model was RF, followed by SVM, LR, and XGBoost. Further, the DCA curves (Figure 2) demonstrate that the RF and SVM models exhibited a greater net benefit along with the threshold probability compared with other models. The accuracy of 9 independent ML models and the stacking model ranked first and second using 10-fold cross-validation are shown in Table S3 in Multimedia Appendix 1. The lack of large discrepancies in each fold validation displayed the good stability of the prediction models.

Table 1. The predictive performance of the 9 independent models.

Model	Features	AUC[a] (95% CI)	Accuracy	Precision	Recall	F score
GNB[b]	9	0.921 (0.916-0.927)	0.859	0.949	0.856	0.900
KNN[c]	15	0.905 (0.830-0.923)	0.884	0.922	0.922	0.922
DT[d]	15	0.905 (0.822-0.922)	0.859	0.903	0.909	0.906
LR[e]	15	0.942 (0.923-0.948)	0.888	0.928	0.922	0.925
SVM[f]	15	0.949 (0.911-0.953)	0.903	0.929	0.941	0.935
RF[g]	15	0.938 (0.914-0.947)	0.922	0.953	0.941	0.947
XGBoost[h]	15	0.941 (0.904-0.945)	0.884	0.911	0.935	0.922
Lightgbm[i]	15	0.940 (0.912-0.945)	0.884	0.923	0.915	0.921
MLP[j]	15	0.951 (0.928-0.957)	0.884	0.939	0.902	0.920

[a]AUC: area under the receiver operating characteristic curve.

[b]GNB: Gaussian Naïve Bayesian.

[c]KNN: k-nearest neighbor.

[d]DT: decision tree.

[e]LR: logistic regression.

[f]SVM: support vector machine.

[g]RF: random forest.

[h]XGBoost: extreme gradient boosting.

[i]Lightgbm: light gradient boosting machine.

[j]MLP: multilayer perceptron.

Figure 2. The ROC and DCA curves of the independent machine learning models. AUC: area under the receiver operating characteristic curve; DCA: decision curve analysis; DT: decision tree; GNB: Gaussian Naïve Bayesian; KNN: k-nearest neighbor; Lightgbm: light gradient boosting machine; LR: logistic regression; MLP: multilayer perceptron; RF: random forest; ROC: receiver operating characteristic; SVM: support vector machine; XGBoost: extreme gradient boosting.
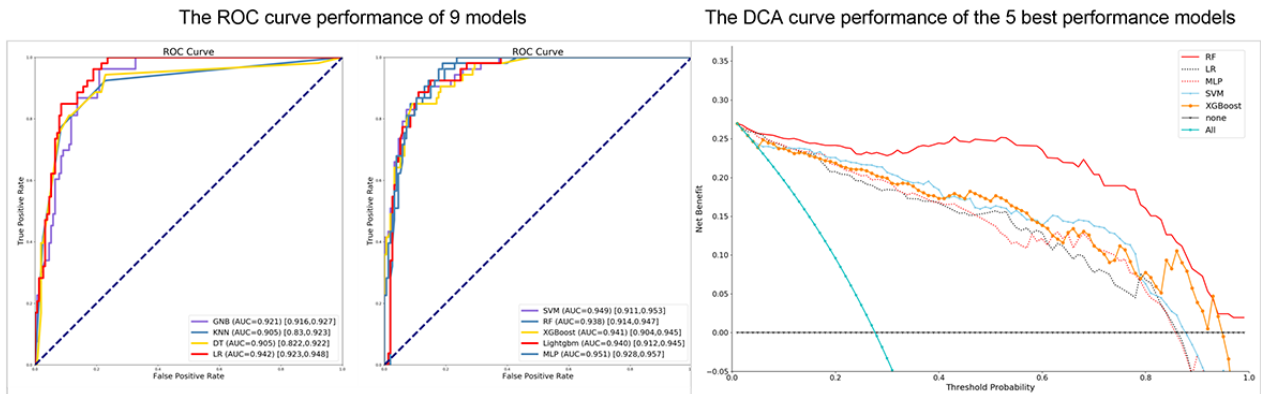
Feature Importance

Figure 3 shows the SHAP plots for SVM and RF. The results reveal that cognitive impairment, care dependency, mobility decline, physical agitation, and an indwelling tube were the strongest predictors. The SHAP plots show that lower levels of these top 5 predictors (ie, blue dots) were associated with a lower probability of PR (ie, SHAP value<0).

Stacking ensemble models were subsequently developed. A total of 510 combinations of different models were output and sorted by CEI, as shown in Multimedia Appendix 2. The stacking ensemble models that ranked first to fourth based on the CEI achieved similar performances in terms of accuracy (0.918), precision (0.942), recall (0.948), and *F* score (0.945). Overall, the first-ranked stacking ensemble model, comprising RF, SVM, and MLP, proved the best under consideration with an AUC value of 0.950 (95% CI 0.924-0.953), which is marginally higher than that of the second-ranked model (AUC 0.949, 95% CI 0.925-0.954; Figure 4). Meanwhile, the stacking model displayed a greater net benefit along with the threshold probability compared with other independent ML models (Figure 4).

Figure 3. Feature importance in the SVM and RF. The SHAP value reflected the impact of features in each sample and performed their positive or negative effects. Contributing factors were ranked in descending order of importance in these plots. Each dot presented a sample; red dots presented a higher feature value, and the right side of the vertical line (ie, feature-specific SHAP values of >0) presented a higher chance of PR use. RF: random forest; SHAP: Shapley Additive Explanations; SVM: support vector machine.
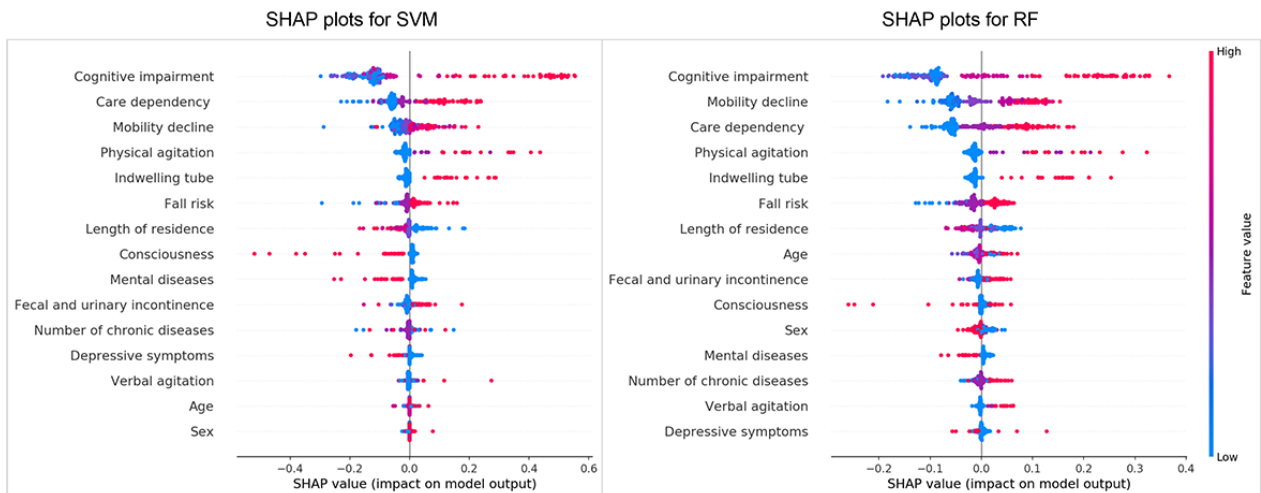
Figure 4. The ROC and DCA curve performance of top 1 stacking ensemble model. AUC: area under the receiver operating characteristic curve; DCA: decision curve analysis; DT: decision tree; MLP: multilayer perceptron; RF: random forest; ROC: receiver operating characteristic; SVM: support vector machine.
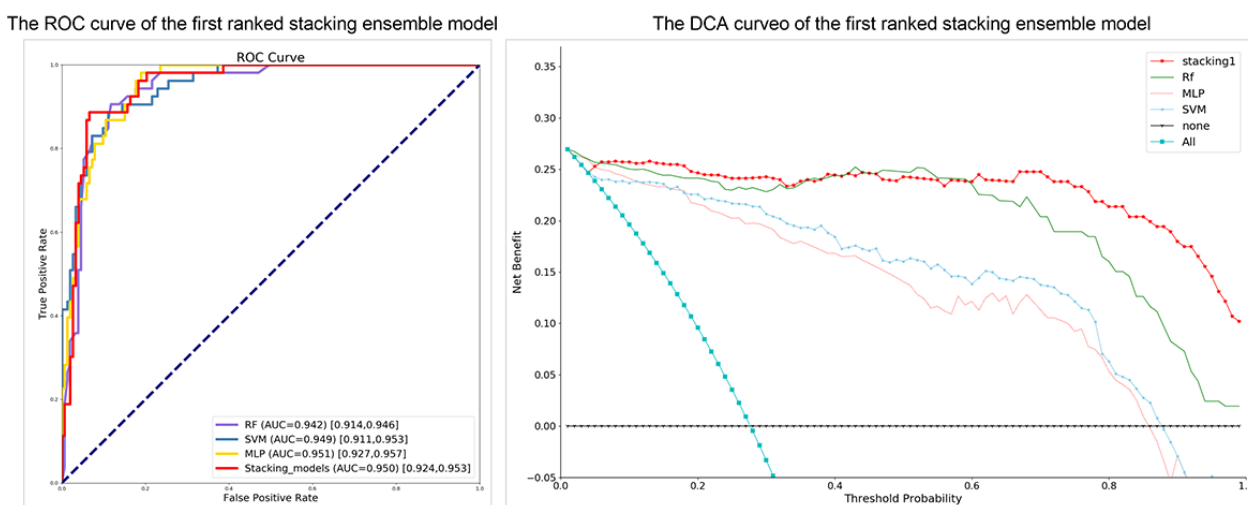
Discussion

Principal Findings

To the best of our knowledge, this is the first study based on the systematic framework of ML techniques to generate multiple models, assess performance, and select the highest-performing models for predicting older adults' possibility of PR use. This study demonstrates that RF and SVM displayed better performance and clinical utility than other independent ML models and that cognitive impairment, care dependency, mobility decline, physical agitation, and an indwelling tube were important contributing factors to model performance. Additionally, the model combining RF, SVM, and MLP was identified as the best model in stacking ensemble learning, which improved the stability, clinical utility, and generalization of the prediction model. These findings contribute to the early identification of older adults at high risk of PR use and targeted clinical care through timely interventions.

Among the independent ML models, SVM and RF were the 2 best-performing models, as identified by the AUC and *F* scores. In small sample data sets, like in this study, SVM may be a good choice for modeling because it can effectively handle high-dimensional spatial data. Moreover, SVM is clearer and more powerful than LR and neural networks when learning complex nonlinear data. The tree-based ML algorithms (eg, RF) are possibly more effective than neural network algorithms in terms of tabular data analysis [45]. This is similar to another study on the PR prediction model in psychiatric patients, in which RF performed better than other algorithms (ie, MLP, LR, and LASSO) [34]. Moreover, Magnowski et al [33] used fast-and-frugal tree modeling to analyze the need for restraint and seclusion; the sensitivity and specificity were only 73% and 76%, respectively. Their model's performance was significantly lower than our model's. The excellent performance of our ML models is primarily due to (1) an iterative grid search procedure for parameter tuning [46], (2) an iterative method for feature selection, and (3) high-quality data with many discrete and without any missing values. Although these models

exhibited good performance and the best one was selected, the independent ML model was inferior to the ensemble learning model in terms of stability and generalization capacity. This limitation was overcome using our stacking-based ensemble learning algorithm [42,47]. Compared with RF, the overall predictive performance (ie, AUC) of the best stacking model combining RF, SVM, and MLP was slightly higher. The levels of sensitivity and precision of the best model are helpful in clinical practice. Additionally, the AHP-stacking algorithm helped to screen optimal models that met the needs of clinical prediction by comprehensively weighing different performance indicators [43].

In a clinical environment, the balance between predictive performance and interpretability must be considered when applying a model. The interpretation of the feature contributions is vital for ML models. Taking the 2 best independent models as examples, we carried out a feature importance analysis and used SHAP plots to visually display how features affected the prediction model. We find that cognitive impairment, mobility decline, and care dependency are the top 3 strongest predictors of PR risk. These features selected in the ML model are consistent with the clinical findings of previous studies [4,5,48]. Conversely, studies have shown that older adults who are restrained have an increased risk of cognitive decline and a decline in ability in activities of daily living [49,50]. These factors interact and result in a vicious cycle. In this study, physical agitation was one of the predictors of PR risk; however, Hofmann et al [51] report no relation between physical agitation and PR risk. Consistent with the clinical situation of using PR to prevent extubation, an indwelling tube increases the risk of PR [6]. This visual interpretation will help highlight important variables for risk prediction and preemptive and early identification of key factors. Subsequently, it will allow nursing staff to develop evidence-based interventions (eg, alternatives to PR) more timely and more targeted and thereby alleviate the risks of the first PR episode. Older adults will experience slower cognitive and ability decline by avoiding PR and engaging in persistent cognitive or ability training, the effect of which is likely to form a virtuous circle and thus reduce the risk of PR

in the future as well. This is an advantage that previous studies on determining risk factors could not achieve. Additionally, as noted in the introduction, little is known about how the interplay of identified risk factors, prediction, and evaluation of PR depends on clinical experience and the decision-makers' subjective judgment without effective and convenient risk assessment tools, which may result in the abuse of PR [52]. These prediction algorithms can automatically calculate the risk for PR without any additional workload. The PR prediction model is intended to be used as a screening tool for predicting potential PR events. An automated PR early warning system, developed based on our results, will offer clinical decision support, which deserves further study in clinical practice. The DCA curves of our models also supported these models' potential clinical utility. Notably, PR prediction is merely a decision-support tool and cannot be relied on for conclusive results. The actual practice should follow the principles of minimized PR and prioritize alternative measures. Previous PR prediction studies of psychiatric inpatients using RF on electronic health data report limitations in adapting other electronic health record systems [34]. In this study, multiple prediction models were developed based on the real risk factors easily collected, which could be helpful in the early triage of older adults and increase the availability of data for the clinical application of the model. Overall, we believe that this work has potential impacts on risk screening, clinical decision-making, and early intervention.

This study had several strengths. First, we adopted a series of widely applied ML algorithms as well as model evaluation techniques that are lacking in existing ML for clinical prediction models. Second, we tuned the hyperparameter values for each ML algorithm identified through an iterative grid search procedure. It has been verified that hyperparameter tuning might improve the performance of models [46]. Most significantly, we presented a stacking framework and implemented it as an ensemble learning algorithm to improve the accuracy and generalization capability of the models [53]. An exhaustive method was used to form 510 combinations to select an appropriate classifier for stacking model construction. For the performance evaluation indicators of the stacking model, the base classifier selection for the stacking algorithm, based on AHP, was adopted in this study. Weight calculations based on

task requirements increase the screening of models that meet clinical demands and practicability.

Limitations

This study has several limitations. Although these models showed excellent performance in internal validation, further external validation using independent database sets is desirable. We collected non-big data from 1026 participants from 6 LTC facilities in 1 city (ie, Chongqing). Clinical large-scale data from various regions of the country should be screened to build a PR prediction model that could be applied widely. Additionally, stacking-based ensemble models were developed without feature screening; however, this had less impact on the performance of the model, as demonstrated by the high-performance results. In the future, the algorithm could be optimized to achieve feature fusion and screening. Furthermore, PR prediction would not depend only on a predictive model, regardless of the significance of the analysis performance. The pragmatic tests of these models in the real world are worth considering. This is an important challenge for all prediction models. Nevertheless, the prediction model of this study has good potential clinical utility in terms of screening assessment tools, clinical decision support, and early intervention, as mentioned in the discussion. User-centered clinical decision support systems or web-based applications based on these models remain a path to better access and improved ease of use [54].

Conclusions

Given the decline in cognitive and daily living functions in older adults, an increasing incidence of PR and adverse effects could be expected. This may impact the quality of care and well-being of older adults. The findings of this study indicate that high-performance ML models for PR risk detection are recommended and have the potential for clinical practice. We identified the high performance and strength of the stacking ensemble learning model in predicting PR use. ML models might facilitate more effective assessments of PR risk and targeted interventions in high-risk individuals. In the future, external validation of multicenter data and the development of a web-based application for better clinical access and ease of use would be worth exploring.

Data Availability

Data can be obtained by contacting the corresponding author (MX).

Authors' Contributions

JW and HC conceptualized and designed the study. WL and DP were responsible for data collection. JW, HC, and HW contributed to the analysis and interpretation of data. JW and HC drafted the paper. WL, QZ, and MX revised the paper. JW, HC, HW, WL, DP, QZ, and MX approved the final version of the manuscript to be submitted.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Descriptions of predictors, risk factors related to physical restraint use identified by univariate analysis, the flow diagram of samples' selection, and accuracy of prediction models using 10-fold cross-validation test.
[DOC File , 173 KB-Multimedia Appendix 1]

Multimedia Appendix 2

The predictive performance of the 510 combinations of different models.
[PDF File (Adobe PDF File), 308 KB-Multimedia Appendix 2]

References

1. Bleijlevens MHC, Wagner LM, Capezuti E, Hamers JPH, International Physical Restraint Workgroup. Physical restraints: consensus of a research definition using a modified Delphi technique. J Am Geriatr Soc 2016;64(11):2307-2310 [FREE Full text] [doi: 10.1111/jgs.14435] [Medline: 27640335]
2. Said AA, Kautz DD. Reducing restraint use for older adults in acute care. Nursing 2013;43(12):59-61. [doi: 10.1097/01.NURSE.0000437484.75992.ca] [Medline: 24257531]
3. Lee DA, Robins LM, Bell JS, Srikanth V, Möhler R, Hill KD, et al. Prevalence and variability in use of physical and chemical restraints in residential aged care facilities: a systematic review and meta-analysis. Int J Nurs Stud 2021;117:103856. [doi: 10.1016/j.ijnurstu.2020.103856] [Medline: 33601305]
4. Huang HC, Huang YT, Lin KC, Kuo YF. Risk factors associated with physical restraints in residential aged care facilities: a community-based epidemiological survey in Taiwan. J Adv Nurs 2014;70(1):130-143. [doi: 10.1111/jan.12176] [Medline: 23734585]
5. Lam K, Kwan JSK, Wai Kwan C, Chong AML, Lai CKY, Lou VWQ, et al. Factors associated with the trend of physical and chemical restraint use among long-term care facility residents in Hong Kong: data from an 11-year observational study. J Am Med Dir Assoc 2017;18(12):1043-1048. [doi: 10.1016/j.jamda.2017.06.018] [Medline: 28843527]
6. Wang J, Liu W, Peng D, Xiao M, Zhao Q. The use of physical restraints in Chinese long-term care facilities and its risk factors: an observational and cross-sectional study. J Adv Nurs 2020;76(10):2597-2609. [doi: 10.1111/jan.14486] [Medline: 33463735]
7. Wolf R, Daichman L, Bennett G. Abuse of the elderly. In: Etienne, Dahlberg LD, editor. World Report on Violence and Health. Geneva: World Health Organization; 2002:123-145.
8. Frank C, Hodgetts G, Puxty J. Safety and efficacy of physical restraints for the elderly. Review of the evidence. Can Fam Physician 1996;42:2402-2409 [FREE Full text] [Medline: 8969858]
9. Perez D, Peters K, Wilkes L, Murphy G. Physical restraints in intensive care-an integrative review. Aust Crit Care 2019;32(2):165-174. [doi: 10.1016/j.aucc.2017.12.089] [Medline: 29559190]
10. Fernández Ibáñez JM, Morales Ballesteros MDC, Montiel Moreno M, Mora Sánchez E, Arias Arias A, Redondo González O. Physical restraint use in relation to falls risk in a nursing home. Rev Esp Geriatr Gerontol 2020;55(1):3-10. [doi: 10.1016/j.regg.2019.05.006] [Medline: 31585682]
11. Sharifi A, Arsalani N, Fallahi-Khoshknab M, Mohammadi-Shahbolaghi F. The principles of physical restraint use for hospitalized elderly people: an integrated literature review. Syst Rev 2021;10(1):129 [FREE Full text] [doi: 10.1186/s13643-021-01676-8] [Medline: 33931096]
12. Bellenger E, Ibrahim JE, Bugeja L, Kennedy B. Physical restraint deaths in a 13-year national cohort of nursing home residents. Age Ageing 2017;46(4):688-693. [doi: 10.1093/ageing/afw246] [Medline: 28049621]
13. Lan SH, Lu LC, Lan SJ, Chen JC, Wu WJ, Chang SP, et al. Educational intervention on physical restraint use in long-term care facilities - systematic review and meta-analysis. Kaohsiung J Med Sci 2017;33(8):411-421 [FREE Full text] [doi: 10.1016/j.kjms.2017.05.012] [Medline: 28811011]
14. Bellenger EN, Ibrahim JE, Lovell JJ, Bugeja L. The nature and extent of physical restraint-related deaths in nursing homes: a systematic review. J Aging Health 2018;30(7):1042-1061. [doi: 10.1177/0898264317704541] [Medline: 28553823]
15. Capeletto CDSG, Santana RF, Souza LMDS, Cassiano KM, Carvalho ACSD, Barros PDFA. Physical restraint in elderly in home care: a cross-sectional study. Rev Gaucha Enferm 2021;42:e20190410 [FREE Full text] [doi: 10.1590/1983-1447.2021.20190410] [Medline: 33886921]
16. Hofmann H, Hahn S. Characteristics of nursing home residents and physical restraint: a systematic literature review. J Clin Nurs 2014;23(21-22):3012-3024. [doi: 10.1111/jocn.12384] [Medline: 24125061]
17. Castle NG, Mor V. Physical restraints in nursing homes: a review of the literature since the nursing home reform act of 1987. Med Care Res Rev 1998;55(2):139-170. [doi: 10.1177/107755879805500201] [Medline: 9615561]
18. Pu H, Zhang X, Luo L, Yu Q, Feng X, Yan L, et al. Use of physical restraints and its associated risk factors in the aged care facilities: a multicentre cross-sectional study. J Clin Nurs 2022 Jun 19:16406. [doi: 10.1111/jocn.16406] [Medline: 35718889]
19. Castle NG. Differences in nursing homes with increasing and decreasing use of physical restraints. Med Care 2000;38(12):1154-1163. [doi: 10.1097/00005650-200012000-00002] [Medline: 11186294]

20. Kor PP, Kwan RYC, Liu JY, Lai C. Knowledge, practice, and attitude of nursing home staff toward the use of physical restraint: have they changed over time? J Nurs Scholarsh 2018;50(5):502-512. [doi: 10.1111/jnu.12415] [Medline: 30058199]

21. Kassew T, Dejen Tilahun A, Liyew B. Nurses' knowledge, attitude, and influencing factors regarding physical restraint use in the intensive care unit: a multicenter cross-sectional study. Crit Care Res Pract 2020;2020:4235683. [doi: 10.1155/2020/4235683] [Medline: 32566288]

22. Ferrão SADS, Bleijlevens MHC, Nogueira PJ, Henriques MAP. A cross-sectional study on nurses' attitudes towards physical restraints use in nursing homes in Portugal. Nurs Open 2021;8(4):1571-1577 [FREE Full text] [doi: 10.1002/nop2.769] [Medline: 33438841]

23. Wang G, Lam KM, Deng Z, Choi KS. Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques. Comput Biol Med 2015;63:124-132. [doi: 10.1016/j.compbiomed.2015.05.015] [Medline: 26073099]

24. Lee SK, Shin JH, Ahn J, Lee JY, Jang DE. Identifying the risk factors associated with nursing home residents' pressure ulcers using machine learning methods. Int J Environ Res Public Health 2021;18(6):2954 [FREE Full text] [doi: 10.3390/ijerph18062954] [Medline: 33805798]

25. Rangrej J, Kaufman S, Wang S, Kerem A, Hirdes J, Hillmer MP, et al. Identifying unexpected deaths in long-term care homes. J Am Med Dir Assoc 2021;23(8):1431.e21-1431.e28. [doi: 10.1016/j.jamda.2021.09.025] [Medline: 34678267]

26. Cary MP, Zhuang F, Draelos RL, Pan W, Amarasekara S, Douthit BJ, et al. Machine learning algorithms to predict mortality and allocate palliative care for older patients with hip fracture. J Am Med Dir Assoc 2021;22(2):291-296 [FREE Full text] [doi: 10.1016/j.jamda.2020.09.025] [Medline: 33132014]

27. Ramazi R, Bowen MEL, Flynn AJ, Beheshti R. Developing acute event risk profiles for older adults with dementia in long-term care using motor behavior clusters derived from deep learning. J Am Med Dir Assoc 2022;23(12):1977-1983.e1. [doi: 10.1016/j.jamda.2022.04.009] [Medline: 35594943]

28. Hu M, Shu X, Yu G, Wu X, Välimäki M, Feng H. A risk prediction model based on machine learning for cognitive impairment among Chinese community-dwelling elderly people with normal cognition: development and validation study. J Med Internet Res 2021;23(2):e20298 [FREE Full text] [doi: 10.2196/20298] [Medline: 33625369]

29. Yang R, Plasek JM, Cummins MR, Sward KA. Predicting falls among community-dwelling older adults: a demonstration of applied machine learning. Comput Inform Nurs 2021;39(5):273-280. [doi: 10.1097/CIN.0000000000000688] [Medline: 33208628]

30. Song W, Kang MJ, Zhang L, Jung W, Song J, Bates DW, et al. Predicting pressure injury using nursing assessment phenotypes and machine learning methods. J Am Med Inform Assoc 2021;28(4):759-765 [FREE Full text] [doi: 10.1093/jamia/ocaa336] [Medline: 33517452]

31. Wong A, Young AT, Liang AS, Gonzales R, Douglas VC, Hadley D. Development and validation of an electronic health record-based machine learning model to estimate delirium risk in newly hospitalized patients without known cognitive impairment. JAMA Netw Open 2018;1(4):e181018 [FREE Full text] [doi: 10.1001/jamanetworkopen.2018.1018] [Medline: 30646095]

32. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319(13):1317-1318. [doi: 10.1001/jama.2017.18391] [Medline: 29532063]

33. Magnowski SR, Kick D, Cook J, Kay B. Algorithmic prediction of restraint and seclusion in an inpatient child and adolescent psychiatric population. J Am Psychiatr Nurses Assoc 2022;28(6):464-473. [doi: 10.1177/1078390320971014] [Medline: 33251913]

34. Danielsen AA, Fenger MHJ, Østergaard SD, Nielbo KL, Mors O. Predicting mechanical restraint of psychiatric inpatients by applying machine learning on electronic health data. Acta Psychiatr Scand 2019;140(2):147-157. [doi: 10.1111/acps.13061] [Medline: 31209866]

35. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ 2015;350:g7594 [FREE Full text] [doi: 10.1136/bmj.g7594] [Medline: 25569120]

36. Laurin D, Voyer P, Verreault R, Durand PJ. Physical restraint use among nursing home residents: a comparison of two data collection methods. BMC Nurs 2004;3(1):5 [FREE Full text] [doi: 10.1186/1472-6955-3-5] [Medline: 15488144]

37. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143(1):29-36. [doi: 10.1148/radiology.143.1.7063747] [Medline: 7063747]

38. Nezu N, Usui Y, Saito A, Shimizu H, Asakage M, Yamakawa N, et al. Machine learning approach for intraocular disease prediction based on aqueous humor immune mediator profiles. Ophthalmology 2021;128(8):1197-1208 [FREE Full text] [doi: 10.1016/j.ophtha.2021.01.019] [Medline: 33484732]

39. Nofallah S, Mehta S, Mercan E, Knezevich S, May CJ, Weaver D, et al. Machine learning techniques for mitoses classification. Comput Med Imaging Graph 2021;87:101832 [FREE Full text] [doi: 10.1016/j.compmedimag.2020.101832] [Medline: 33302246]

40. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. 2017 Presented at: 31st Conference on Neural Information Processing Systems; December 4-9, 2017; Long Beach, CA.

41. Kwon H, Park J, Lee Y. Stacking ensemble technique for classifying breast cancer. Healthc Inform Res 2019;25(4):283-288 [FREE Full text] [doi: 10.4258/hir.2019.25.4.283] [Medline: 31777671]

XSL•FO

RenderX

42. Xiong Y, Ye M, Wu C. Cancer classification with a cost-sensitive Naive Bayes stacking ensemble. Comput Math Methods Med 2021;2021:5556992 [FREE Full text] [doi: 10.1155/2021/5556992] [Medline: 33986823]

43. Sun T, Chen Y. Base classifiers selection for stacking algorithm based on AHP. Ordnance Industry Automation 2022(1):39-42 [FREE Full text] [doi: 10.7690/bgzdh.2022.01.009]

44. Schmidt K, Aumann I, Hollander I, Damm K, von der Schulenburg JMG. Applying the analytic hierarchy process in healthcare research: a systematic literature review and evaluation of reporting. BMC Med Inform Decis Mak 2015;15:112 [FREE Full text] [doi: 10.1186/s12911-015-0234-7] [Medline: 26703458]

45. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of machine learning models to predict death after acute myocardial infarction. JAMA Cardiol 2021;6(6):633-641 [FREE Full text] [doi: 10.1001/jamacardio.2021.0122] [Medline: 33688915]

46. Wong J, Manderson T, Abrahamowicz M, Buckeridge DL, Tamblyn R. Can hyperparameter tuning improve the performance of a super learner?: a case study. Epidemiology 2019;30(4):521-531 [FREE Full text] [doi: 10.1097/EDE.0000000000001027] [Medline: 30985529]

47. Han S, Wang Y, Liao W, Duan X, Guo J, Yu Y, et al. The distinguishing intrinsic brain circuitry in treatment-naïve first-episode schizophrenia: ensemble learning classification. Neurocomputing 2019;365:44-53. [doi: 10.1016/j.neucom.2019.07.061]

48. Estévez-Guerra GJ, Fariña-López E, Núñez-González E, Gandoy-Crego M, Calvo-Francés F, Capezuti EA. The use of physical restraints in long-term care in Spain: a multi-center cross-sectional study. BMC Geriatr 2017;17(1):29 [FREE Full text] [doi: 10.1186/s12877-017-0421-8] [Medline: 28109267]

49. Freeman S, Spirgiene L, Martin-Khan M, Hirdes JP. Relationship between restraint use, engagement in social activity, and decline in cognitive status among residents newly admitted to long-term care facilities. Geriatr Gerontol Int 2017;17(2):246-255. [doi: 10.1111/ggi.12707] [Medline: 26822624]

50. Sze TW, Leng CY, Lin SKS. The effectiveness of physical restraints in reducing falls among adults in acute care hospitals and nursing homes: a systematic review. JBI Libr Syst Rev 2012;10(5):307-351. [doi: 10.11124/jbisrir-2012-4] [Medline: 27819940]

51. Hofmann H, Schorro E, Haastert B, Meyer G. Use of physical restraints in nursing homes: a multicentre cross-sectional study. BMC Geriatr 2015;15:129 [FREE Full text] [doi: 10.1186/s12877-015-0125-x] [Medline: 26489664]

52. Wang J, Liu W, Zhao Q, Xiao M, Peng D. An application of the theory of planned behavior to predict the intention and practice of nursing staff toward physical restraint use in long-term care facilities: structural equation modeling. Psychol Res Behav Manag 2021;14:275-287 [FREE Full text] [doi: 10.2147/PRBM.S293759] [Medline: 33688280]

53. Liang X, Li F, Chen J, Li J, Wu H, Li S, et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. Brief Bioinform 2021;22(4):bbaa312 [FREE Full text] [doi: 10.1093/bib/bbaa312] [Medline: 33316035]

54. Kim HJ, Han D, Kim JH, Kim D, Ha B, Seog W, et al. An easy-to-use machine learning model to predict the prognosis of patients with COVID-19: retrospective cohort study. J Med Internet Res 2020;22(11):e24225 [FREE Full text] [doi: 10.2196/24225] [Medline: 33108316]

Abbreviations

AHP: analytic hierarchy process
AUC: area under the receiver operating characteristic curve
CEI: comprehensive evaluation indicator
DCA: decision curve analysis
DT: decision tree
GNB: Gaussian Naïve Bayesian
KNN: k-nearest neighbor
LR: logistic regression
LTC: long-term care
ML: machine learning
MLP: multilayer perceptron
PR: physical restraint
RF: random forest
SHAP: Shapley Additive Explanations
SVM: support vector machine
XGBoost: extreme gradient boosting