Development and Validation of Interpretable Machine Learning Models for Inpatient Fall Events and EMR Integration

Abbreviated title: Interpretable machine learning for Inpatient fall

Authors: Soyun Shim 1†, Jae Yong Yu 2,3†, Seyong Jekal 4, Yee Jun Song 2, Ki Tae Moon 4, Ju Hee Lee 1, Kyung Mi Yeom 1, Sook Hyun Park 1, In Sook Cho 5, Mi Ra Song 1, Won Chul Cha 2,4,6, and Jung Hee Hong 1*

Affiliations: Full information for each author

1 Department of Nursing, Samsung Medical Center, 115 Irwon-ro Gangnam-gu, Seoul 06355, Republic of Korea.
2 Department of Digital Health, Samsung Advanced Institute for Health Science & Technology (SAIHST), Sungkyunkwan University, 115 Irwon-ro Gangnam-gu, Seoul 06355, Republic of Korea.
3 Digital and Smart Health Office, Tan Tock Seng Hospital (TTSH), Singapore
4 Digital Innovation Center, Samsung Medical Center, 115 Irwon-ro Gangnam-gu, Seoul 06355, Republic of Korea
5 Department of Nursing, In-ha University, Incheon, Republic of Korea.
6 Department of Emergency Medicine, Samsung Medical Center, Sungkyunkwan University School of Medicine, 115 Irwon-ro Gangnam-gu, Seoul 06355, Republic of Korea.

†Soyun Shim and Jae Yong Yu contributed equally to this work.

Address for Correspondence:

* Corresponding author: Jeong Hee Hong
Department of Nursing, Samsung Medical Center, 115 Irwon-ro Gangnam-gu, Seoul 06355, Republic of Korea.
Tel: +82-2-3410-2900; Fax: +82-2-3410-0004; E-mail: jhee.hong@samsung.com
Capsule Summary

What is already known: Conventional score such as Morse fall scales was developed for fall risk assessment, however it has low accuracy and difficulty in practical implementation. Previous machine learning studies have limitation in its explainability for clinicians and nurses, which is also called as the black box problem. Another challenge is that previous models have not been well-prepared for Electronic Medical Records integration for practical application.

What is new in the current study:

We developed the interpretable Machine learning model for the prediction of falls by integrating into EMR system.
ABSTRACT

(1) **Background:** Falls are one of the most frequently occurring adverse events among hospitalized patients. The Morse fall scales, which have been widely used for fall risk assessment, has two limitations: low specificity and difficulty in practical implementation. The aim of this study was to develop and validate an interpretable machine learning (ML) model for the prediction of falls, which would be integrated in an electronic medical record system (EMR).

(2) **Methods:** This is a retrospective study from a tertiary teaching hospital in Seoul, Korea. Data were collected from January 2018 to March 2020. Based on literatures, known 83 predictors were grouped into seven categories. Interpretable fall event prediction models were developed using multiple machine learning models including gradient boosting and Shapley values.

(3) **Results:** Overall, 191,778 cases with 272 fall events (0.1%) were included for the analysis. With the validation cohort of 2020, the area under the receiver operating curve of the gradient boosting model was 0.817 (95% confidence interval, 0.720 – 0.904) which showed better performance than random forest, logistic regression, artificial neural net, and conventional Morse fall score (0.80, 0.80, 0.74, and 0.65, respectively). The model’s interpretability was enhanced in both population level and patient level. The algorithm was later integrated into the current EMR.

(4) **Conclusion:** We developed an interpretable ML prediction model for inpatient fall events using EMR integration formats.

**Keywords:** Inpatient Falls, Adverse Event, Machine Learning, Prediction Model, Patient Safety, Nursing Informatics
1. Introduction

Falls are one of the most frequently occurring adverse events in hospitalized patients, extending the hospital stay, increasing medical costs, and increasing disability and mortality. In the United States, falls occur at a rate of 3.3 to 11.5 per 1,000 hospitalized days. Among patient safety accidents in Korea, falls were the most reported over the past 5 years, accounting for 45% or more, and more than two-thirds of them resulted in mild or more serious injuries.

Nurses must assess patients’ fall risk and, if necessary, provide appropriate caution and education. According to the Joint Commission on Healthcare Organization Accreditation and the Korea Institute for Healthcare Accreditation, falls are critical incidents, that preventing falls must be highly prioritized as a hospital policy. The Morse fall scale (MFS) is the most widely used fall risk assessment scales along with other tools.

The MFS focuses primarily on intrinsic factors which are related to the individual factor such as history of falls or polypharmacy and poses major limitations in two aspects. First, the prediction accuracy of MFS varies significantly among different healthcare settings and patient groups making its application challenging due to the need for real-time intervention to response for changing patient conditions. Second, because the MFS does not focus on individual risk factors, nurses must apply a broad fall prevention plan without necessarily focusing on individual patient's unique risk factors.

Advances in data science have contributed to the development of accurate fall risk prediction models other studies created ensemble model or extreme gradient boosting model and identified significant predictors such as low self-care ability, sleep disorder and medication use. Though with acceptable accuracy, previous machine learning (ML) studies have limitation in its explainability for clinicians and nurses, which is also called as the black box problem. Another challenge is that previous models have not been well-
prepared for Electronic Medical Records (EMR) integration for practical application \textsuperscript{15,16}

To solve these limitations, we developed the model for prediction of falls using interpretable ML and integrated model into EMR system to perform nursing interventions for each risk factor.
2. Methods

2.1. Study Setting

This retrospective study was conducted in a tertiary academic hospital in Seoul, Korea. The hospital had approximately 2,000 inpatient beds. Patient’s data from six general medical surgical wards were involved for data collection considered where falls were reported to be high. Data were obtained from EMR. This study was approved by the hospital’s institutional review board (IRB) (No. SMC 2022-03-052-001). Because of the retrospective nature of the study, participants’ consent was waived. The TRIPOD (transparent reporting of a multivariable prediction model for individual prognosis or diagnosis) statement was followed for the development and reporting of multivariable prediction models 17.

2.2. Study Population

All patients who were admitted to the six general wards from January 2018 to March 2020 were included in the study. Patients aged < 18 years, those who had a length of stay of < 24 hours, and those with multiple fall injuries during the same admission were excluded. If patients admitted more than once, each admission was evaluated independently. These admissions were split into two non-overlapping cohorts for temporal validation: a development cohort from January 2018 to December 2019 and a testing cohort from January 2020 to March 2020 for the evaluation of the model.

2.3 Candidate Predictors

Eighty-three candidate predictors were originally suggested based on a conceptual model of inpatient fall risk concepts from multiple national fall prevention guidelines in the United States and Korea 18. Demographic and admission data, physician orders, and nurse records were selected as potential predictors. These variables were collected from patient data in the
EMR via flowsheets, medical diagnoses, nursing care plans, and free text notes. In previous studies, nurse intervention for fall prevention was reported to be an important variable in the occurrence of falls, so nursing care plan was used in this study. Here, as shown in Figure 1, each candidate predictor was defined using within 24 hours from admission with 1 hour of time frame for predicting the falls every hour.

Each circle means each candidate predictor and the number of circles means how many times of each predictor were measured within 24 hours. Our first prediction result can be calculated 24 hours later from admission, so we defined time $t$ as prediction index time which is admission time + 24.

For practical application for clinical providers, these 83 candidate predictors were mapped into seven categories: universal, cognitive function, defecation problem, mobility problem, medication, sensory function, and sleep disturbance, according to “Evidence based clinical nursing practice guideline of Korea Hospital Nurses Association”. The list of predictors with known clinical and statistical significance, therefore included in the model, is presented in Table 1. There were nursing assessment and intervention that nurses frequently recorded for each shift, and the records of initial nursing evaluation at the time of admission were used. In addition, the KPCS score, which is input to inpatients every day, was also used. The predictor with the highest contribution was nursing intervention.

### 2.4. Outcome of prediction models

Two distinct sources were defined as the primary outcomes for predicting falls. One is based on the existence of an incident reporting system; The other is based on a regular expression of a total of 11 fall-related terms, such as "fallen" and "slipped", from a nursing free text record. Fall events were manually cross-checked by two nurses.
2.5. Statistical Analysis

For statistical analysis, Python version 3.6.0 and SQL version 3.6.6 were used.

Continuous variables were described as means and standard deviations (SD). Categorical variables were described as frequencies and percentages. The t-test and chi-square test were used to calculate the p-value where p<0.05 was considered statistically significant.

The development cohort was used to develop the prediction model, and the testing cohort was used to optimize the hyper-parameter for it. Then, the performance metrics of the final model were calculated based on the testing cohort.

During the analysis, the missing value for KPCS score; fluid management were imputed with recent non missing value and 0 imputation for the other count values.

2.6. Machine Learning Models

Adapted from a previous pilot study, which used the Bayesian network model derived from January 2017 to June 2018 and showed 0.93 of area under the receiver operating characteristic (AUROC) performance; a XGB (eXtreme Gradient Boosting) model was developed, which is a fast and scalable ML technique for tree-based ensemble models.

Hyper-parameter tuning for XGB was conducted by considering the grid search of maximum depth, number of estimators, learning rate, etc., with the highest AUROC performance in the 10-fold cross-validation set.

The AUROC and area under the precision-recall curve (AUPRC) in the validation datasets were calculated. To obtain 95% confidence intervals (CIs), 500 bootstrap repetitions were
conducted. The sensitivity, specificity were calculated using the Youden index, which is defined as the point nearest to the upper-left corner of the receiver operating characteristic (ROC) curve.

The prediction models were compared with the conventional point-based MFS, which consists of six evaluation items including a history of falling, secondary disease, ambulatory aid, intravenous therapy/heparin lock, gait, and mental status with and without nurses’ judgment. To supplement the MFS, the hospital identified high-risk individuals based on clinical judgments made by the nurse.

Furthermore, other traditional ML methods were used for comparison. L2 regularized logistic regression, random forest, and essential artificial neural network (ANN) (three layers) were performed with default settings. The software implemented for model development and validation were Python programming language (version 3.8.5), TensorFlow framework (version 2.3.1), and scikit-learn (version 0.23.2).

2.7. Model explainability for EMR integration

To apply the ML model to the clinical environment, the same model structure was developed for the EMR integration. Seven categorized feature contributions to falls were identified at the patient level using the Shapley. The Shapley value has been widely used in game theory literature for calculating the contribution of each player in the game; and in terms of the prediction modeling area, the Shapley value can be used to compute the contribution of each data point to the model’s final performance and can be visualized with shap force plot as shown in Figure 2. The ML model, which suggests patient risk factors according to the Shapley value, was integrated into EMR and screen development was implemented so that it could be used in actual clinical settings.
Two components, the predictive probability of falls and the predictive risk factor were provided in EMR screen. According to the threshold, it was classified as high risk or low risk and the top three categorized feature contributions were selected for the preceding intervention. If the result is classified as a low risk and the nurse does not agree with the result, it is allowed to directly change to a high risk. If the nurse agrees on the predictive result and the risk factors, the recommended intervention is performed and recorded. When a low-risk patient becomes a high-risk patient, a pop-up screen informs the nurse of the status change.
3. Results

3.1. Basic characteristics

Initially, 257,140 cases between January 2018 and March 2020 were included. A total of 65,362 cases were excluded for cases under 18 years of age, undefined general words, events on the same admission day, or multiple falls during the same admission. A total of 191,778 cases (272 cases (0.14%) in the group with falls) were used for the final data analysis. The flow diagram of the study process is shown in Figure 3.

Their basic characteristics are listed in Table 1. During the study period, 272 (0.14%) fall cases (mean age (SD), 62.08 (14.63) years; 167 (61.40%) male) were reported. There were more male and elderly patients in the fall group. The fall group was taking more medication was more likely to take medication and have a secondary diagnosis.

3.2. Machine Learning Models

An ML-based fall prediction algorithm was developed. Table 2 summarizes the AUROC, AUPRC, and other metrics using various methods with 95% CI. The best AUROC and AUPRC was 0.82 (0.72 – 0.90) and 0.01 (0.01 – 0.02) for gradient boost model (GBM). The AUROC plot is shown in Figure 4.

According to the cut-offs determined by the Youden index, the sensitivities (95% CI) were 0.81 (0.81 – 0.82), 0.54 (0.35 – 0.74), 0.71 (0.50 – 0.89), 0.67 (0.50 – 0.85), 0.83 (0.68 – 0.96) and 0.96 (0.87 – 1.00), and the specificities (95% CI) were 0.81 (0.81 – 0.82), 0.91 (0.90 – 0.91), 0.74 (0.73 – 0.74), 0.64 (0.63 – 0.65), 0.47 (0.46 – 0.48) and 0.33 (0.32 – 0.34) respectively.

The contribution of the event to seven categorized factors was identified in terms of population and patients, as shown in Figure 5. The sum of medication taken, number of
nursing interventions for fall prevention, and education were the most influential factors in predicting falls for the population level.

Regarding the patient level prediction, as shown in Figure 5 (B), a patient’s risk was demonstrated as individual factors using Shapley value. The patient-level predictors are displayed in the EMR (Figure 6). The reason of prediction was displayed as a binary output (check-box) by groups: cognitive function, toileting problem, mobility problem, medication, sensory function, sleep disturbance.
4. Discussion

In this study, an interpretable ML for fall event prediction was developed using 83 predictors from the EMR data, with the gradient booting model demonstrating the highest performance. By presenting the risk factors, predictors were categorized as fall risk factors and incorporated into the EMR. This can be the cornerstone for the development of an AI-based clinical decision support system (CDSS) in the hospital regarding the application to the clinical field.

Patients in this study who experienced falls tended to be older, stayed in the hospital longer, and took more medications. This result is consistent with the characteristics of patients at risk for falls, as in previous studies\(^\text{14, 18}\). All these characteristics were used as significant predictors.

The model performance of this study was AUROC 0.817, which was higher than that of previous studies\(^\text{13, 14}\). Medication and nursing intervention was important predictors consistent with previous study\(^\text{18}\). Lower limb weakness and dysuria was the highest predictor in previous study\(^\text{14, 22}\), assist toileting and ambulation nursing interventions were found to be high predictor, which was similarly consistent. Another study shows admission data was high in feature importance, but this study inputted more variables to reflect specific patient conditions.\(^\text{14}\). Our study developed a model by including not only formally reported falls but also falls that were only recorded in nurses’ clinical notes. Data that reflects patient status at the time of a fall event was used based on previous literatures\(^\text{22}\). With all these attempts, we could improve the model’s sensitivity and specificity not just as a model’s accuracy, but also as accuracy in clinical application setting.

Interpretable ML is critical for nursing practice application\(^\text{24, 25}\). Accuracy and actionable intervention is important for falls case because falls are very sensitive event that might result
in hospital lawsuits. The findings of this study will enable the identification of risk factors that can guide individualized interventions.

Although many ML algorithms have been developed to predict falls, pressure injuries, and delirium \textsuperscript{13, 23, 24}, to our best knowledge, there was no study which is integrated and applied to actual EMR. Prospective studies could be evaluated to determine the applicability and usefulness of AI-based CDSS for the future study.

AI for patient safety is a very impactful and effective area. It can be extended as a valuable tool that can be used to improve patient safety in multiple clinical settings: healthcare includes healthcare-associated infections, adverse drug events, venous thromboembolism, surgical complications, pressure ulcers, falls, decompensation, and diagnostic errors \textsuperscript{25}.

This study had some limitations. First, the study was performed in a single center in a retrospective manner. To evaluate the model performance, further large clinical datasets and multicenter validation are required. Prospective studies could be evaluated to verify the algorithm performance with the CDSS, which involve evaluation of its effectiveness and usability in work process and safety outcome.

Second, in terms of predictors, the dataset used in this study was based on routinely collected EMR variables. Thus, other factors as environmental and behavioral ones, which are important but merely recorded in the EMR could not be utilized.

Finally, the outcome for the fall was only collected as a EMR reports. Near-miss cases as stumbling and sliding which had not been reported or recognized by providers have potential importance that further studies should involve sensors and reports by patients to include these cases.

In conclusion, we developed an interpretable ML prediction model for fall events, which were integrated into EMR. This is one of the first attempts to integrate AI-CDS into practice
in a large scale, which will require further studies regarding its effectiveness and safety.
Acknowledgements: The authors thank all the participants in the expert interview and survey for their assistance during this study.

Funding: This research did not receive any specific grants from funding agencies in the public, commercial or not-for-profit sectors.

Author contributions: Conceptualization: WC Cha; Data curation: SY Chegal, YJ Song, KT Moon; Formal analysis: SY Chegal, YJ Song, KT Moon; Investigation: SY Shim, JY Yu, JH Lee, KM Yeom, SH Park, IS Cho, MR Song, JH Hong; Methodology: SY Chegal, YJ Song, KT Moon; Visualization: SY Shim, JY Yu, SY Chegal, YJ Song, KT Moon; Writing - original draft: SY Shim, JY Yu, WC Cha, Writing - review & editing: SY Shim, JY Yu, WC Cha, JH Hong

Institutional Review Board Statement: “The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board of Samsung Medical Center (SMC 2022-03-052-001).”

Conflicts of Interest: “The authors declare no conflict of interest.”

References


13. Hsu YC, Weng HH, Kuo CY, Chu TP, Tsai YH. Prediction of fall events during admission


23. O'Brien RL, O'Brien MW. CE: Nursing Orientation to Data Science and Machine


**Table 1.** Detailed list of candidates and known clinically significant predictors and seven categorized classes.

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>Predictors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universal</td>
<td>25</td>
<td>Age, Sex, Primary and secondary medical diagnosis, Medical department, History of falls, Length of stay, KPCSa, Numbers of medication, Move-in date, Dates of surgical operation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nursing assessment: Mental status, RASSb, Nursing diagnosis: Acute Chronic confusion, Nursing intervention: Provide bed alarm with bed sensor pad, Restraint</td>
</tr>
<tr>
<td>Cognitive function</td>
<td>11</td>
<td>Nursing intervention: Timed voiding, Provide portable toilet seat, Medication (Diuretics, Laxative), Nursing assessment: Urine output, Stool count, Nursing diagnosis: Impaired urination, Diarrhea</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nursing intervention: Provide assistive device, Fluid management, Nursing assessment: Dizziness, ADLc, Aid, Deformity, Disability, Nursing diagnosis: Impaired mobility, Fall risk medication (Sedatives, Antidepressant, Antiemetics, Antipsychotics, Antianxiety drugs, Antihypertensives, Analgesics, Antiarrhythmics and NSAIDsd), Nursing assessment: Catheter (Central venous line and intravenous line), Adverse drug reaction monitoring</td>
</tr>
<tr>
<td>Toileting problem</td>
<td>7</td>
<td>Nursing assessment: Sensory, Motor, Circulation, Nursing diagnosis: Sensory perception, Nursing intervention: Provide assistive device</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nursing assessment: Sleep pattern, Delirium, Nursing diagnosis: Disturbed sleep pattern, Nursing intervention: Sleep enhancement, Medication (Antianxiety drugs)</td>
</tr>
<tr>
<td>Mobility problem</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>Medication</td>
<td>10</td>
<td>KPCS : Korean Patient Classification System b)RASS : Richmond Agitation Sedation Scale c)ADL : Activities of Daily Living d)NSAIDs : non-steroidal anti-inflammatory drug; known clinically significant variables were marked in italic</td>
</tr>
</tbody>
</table>
Table 2. Basic characteristics of the study population.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Falls (N=272)</th>
<th>No Falls (N=191,506)</th>
<th>P-value(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, mean (SD(^a))</td>
<td>62.1 (14.6)</td>
<td>59.8 (14.2)</td>
<td>0.012</td>
</tr>
<tr>
<td>Gender, n (%)</td>
<td></td>
<td></td>
<td>0.944</td>
</tr>
<tr>
<td>- Male</td>
<td>167 (61.4)</td>
<td>116,830 (61.0)</td>
<td></td>
</tr>
<tr>
<td>- Female</td>
<td>105 (38.6)</td>
<td>74,676 (39.0)</td>
<td></td>
</tr>
<tr>
<td>Length of Stay, mean (SD)</td>
<td>18.9 (21.6)</td>
<td>23.5 (160.5)</td>
<td>0.001</td>
</tr>
<tr>
<td>KPCS(^c), n(%)</td>
<td></td>
<td></td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>- group 1 (less severe)</td>
<td>46 (17.0)</td>
<td>43,676 (25.8)</td>
<td></td>
</tr>
<tr>
<td>- group 2</td>
<td>165 (60.9)</td>
<td>82,514 (48.7)</td>
<td></td>
</tr>
<tr>
<td>- group 3</td>
<td>48 (17.7)</td>
<td>31,886 (18.8)</td>
<td></td>
</tr>
<tr>
<td>- group 4 (most severe)</td>
<td>12 (4.4)</td>
<td>11,235 (6.7)</td>
<td></td>
</tr>
<tr>
<td>Daily medication per person (Past 4 weeks), mean (SD)</td>
<td>21.3 (28.3)</td>
<td>15.9 (37.9)</td>
<td>0.002</td>
</tr>
<tr>
<td>Patients Classification, n(%)</td>
<td></td>
<td></td>
<td>0.219</td>
</tr>
<tr>
<td>- Surgical</td>
<td>83 (30.5)</td>
<td>63,873 (33.4)</td>
<td></td>
</tr>
<tr>
<td>- Medical</td>
<td>189 (69.5)</td>
<td>127,633 (66.6)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\) SD : standard deviation; \(^c\) KPCS : Korean Patient Classification System
\(^b\) p-value were calculated with t-test for continuous predictors and chi-square test for categorical predictors.
Table 3. Comparison of evaluation values with 95% Confidence Interval achieved by different methods on the Testing Cohort.

<table>
<thead>
<tr>
<th>Method</th>
<th>AUROC</th>
<th>AUPRC</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>0.817</td>
<td>0.010</td>
<td>0.750</td>
<td>0.811</td>
</tr>
<tr>
<td>Boost</td>
<td>0.801</td>
<td>0.010</td>
<td>0.542</td>
<td>0.907</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.798</td>
<td>0.005</td>
<td>0.350</td>
<td>0.903</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.802</td>
<td>0.011</td>
<td>0.708</td>
<td>0.736</td>
</tr>
<tr>
<td>Regression</td>
<td>0.736</td>
<td>0.008</td>
<td>0.750</td>
<td>0.640</td>
</tr>
<tr>
<td>ANN</td>
<td>0.650</td>
<td>0.002</td>
<td>0.583</td>
<td>0.633</td>
</tr>
<tr>
<td>MFS</td>
<td>0.645</td>
<td>0.002</td>
<td>0.833</td>
<td>0.470</td>
</tr>
<tr>
<td>MFS with Judge</td>
<td>0.598</td>
<td>0.001</td>
<td>0.867</td>
<td>0.324</td>
</tr>
</tbody>
</table>

a) AUROC, area under the receiver operating characteristic; b) AUPRC, area under the precision-recall curve; c) ANN, artificial neural network; d) MFS, Morse Fall Scale.
Figure 1. Dataset generation for each variable from one admission. The time period was defined as 24-hours from the admission time and 1 hour time window.
Figure 2. Shap force plot for visualization of individualized prediction result by risk factor for fall prediction.
Figure 3. Flowchart of the study
Figure 4. Comparison of receiver operating characteristic (ROC) curves for various methods; GBoost, eXtreme Gradient Boosting, ANN, artificial neural network MFS, Morse Fall Scale.
Figure 5. Population and patient-level interpretation for fall events. (A) Population-level interpretation with feature importance by gradient Boosting (B) Patient-level interpretation with Shapley value.
Figure 6. Clinical Decision Support System of fall prediction model in the Electronic Medical Records. In the bold yellow box, the patient-level top two contributing factors are automatically checked (red check) from six categorized predictors.