Predicting Efficacy of Rehabilitation for Orthopedic Patients Using Machine Learning

Caleb Oh, Douglass Johnson, Addisu Mesfin
Georgetown University, Washington DC
co529@georgetown.edu

Disclosures: Caleb Oh (N), Douglass Johnson (N), Addisu Mesfin (N)

INTRODUCTION: Following an orthopedic procedure, an important question for both patients and physicians is when and to what extent a patient will be able to recovery daily functions. This study aims to explore the feasibility of employing machine learning models to predict levels of improvement that orthopedic patients can expect through rehabilitation following a procedure. A publicly available dataset of 1844 patients hospitalized from 2015 to 2018 after hip prosthesis, knee prosthesis, femur internal fixation, fraction of spinae and trunk, amputation, or other orthopedic procedures at the "San Raffaele" rehabilitation hospital was downloaded and used for data exploration, data cleaning, feature engineering, and generation of machine learning models. In this study, six machine learning models were trained using 75 patient characteristics to predict level of improvement— 'no change', 'minimal improvement', 'moderate improvement', or 'significant improvement'— that each patient would experience following rehabilitation.

METHODS: The dataset used, which was made publicly available by Seccia et al. in 2020, was downloaded as a csv file containing 1844 unique patients and 75 features (1844 x 75). We performed all data manipulation, modeling, and data visualization using a Jupyter Notebook and Python packages NumPy, scikit-learn, Matplotlib, and Seaborn. After exploring the 75 features in the dataset, we addressed columns with missing values, renamed all the column headers according to their full description, and created a new categorical outcome column titled 'improvement'. We designed this column to calculate the difference in Barthel Index (a measure of daily functional ability) from the time of admission to the time of discharge for each patient. Then we create thresholds based on these differences (Δ) in Barthel Index and binned them into the following categories: (1) significant improvement: $\Delta > 20$; (2) moderate improvement: $10 < \Delta <= 20$; (3) minimal improvement: $1<= \Delta <=10$; no change: $\Delta = 0$. This new column titled 'improvement' became our outcome of interest that we would try to predict given 75 patient features. Once the data engineering was complete, various machine learning models were employed, including neural networks, SVM, KNN, logistic regression, random forest, and gradient boost, trained on a 70-30 split dataset. The performance was evaluated using macro F1, weighted F1 scores, and overall accuracy of predictions.

RESULTS SECTION: Of the six models, the highest performing classifier was Support Vector Machine (SVM) which produced an overall accuracy of 98.56%, macro average of 68% and weighted average of 99%. The overall accuracy based on precision, recall, f1-score, and support for predictions made in each improvement category was over 94% accurate for all models. The performance of all the machine learning models is compared in Figure 1. The confusion matrix for the predictions of the SVM is displayed in Figure 2. A chart of the most important features informing predictions for the SVM model is displayed in Figure 3 which include chair/bed transfer, dressing, bladder control, toilet use, personal hygiene at time of admission among others.

DISCUSSION: The literature on using machine learning to accurately predict the efficacy of rehabilitation is sparse, especially for orthopedic specific pathologies. This can be attributed, in part, to the lack of high-quality datasets that are made publicly available. However, with the increasing trend of 'dividing the labor' where one study focuses on producing a high-quality dataset and encouraging other investigators to run the models (Seccia et al., 2020), the ability to create highly accurate models becomes more feasible. For the original dataset, based on our knowledge, only one other study has created machine learning models using the same data, specifically employing 4 ensemble models to predict the modified Barthel index at the time of discharge (Santilli et al., 2023). In our study, we decided not to use a continuous outcome variable like the modified Barthel index at time of discharge but instead created a new outcome variable that we titled 'improvement'. The reasoning for this is that although the modified Barthel index provides great insight into functional ability post-rehabilitation, this approach precludes the starting point of patients upon admission and the distance they have gone to achieve their Barthel score at discharge. Another reason for modifying our outcome variable was that we wanted to use a categorical outcome, which may be more easily understood by patients and physicians compared to a continuous variable like the Barthel index score which ranges from 0-100. That is, it may be harder to interpret number than a specific description like 'significant improvement'. In terms modeling, a wide variety of machine learning models were tested from logistic regressions to ensemble methods to Neural networks. The data was split 70-30 into a training set with 1290 patients in the train set and 554 patients in the test set. Limitation for this study include the retrospective nature of the dataset and imbalanced outcome categories in favor of patients who have had significant improvement. However, to account for the imbalance, both the macro average and weighted average were reported.

SIGNIFICANCE/CLINICAL RELEVANCE: The ability to use machine learning models to predict the individualized improvement that orthopedic patients can expect from rehabilitation may provide deeper insight to both patients and physicians. Furthermore, high performing model (SVM = 98%) can help clarify which variables or signs are most predictive of improvement during rehabilitation.

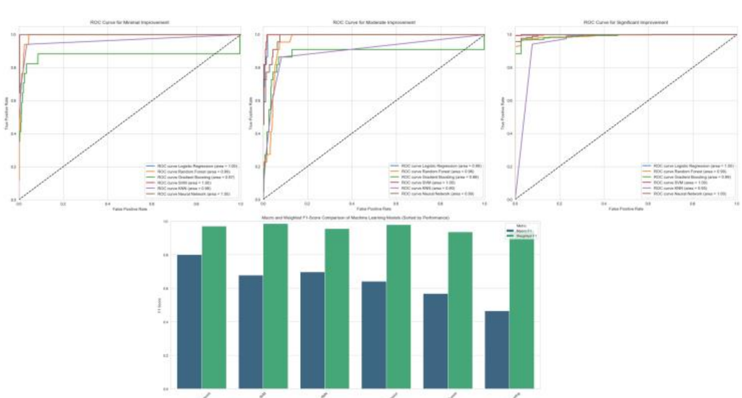
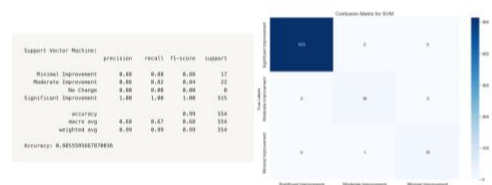


Figure 1: The ROC curves on the top row compare the performance of each machine learning model for each classification task. The row on the bar chart on the bottom row show the macro F1 and weighted F1 scores for each model side-by-side.



Support Vector Machine:

	precision	recall	f1-score	support
Minimal Improvement	0.88	0.88	0.88	17
Moderate Improvement	0.86	0.82	0.84	22
No Change	0.00	0.00	0.00	0
Significant Improvement	1.00	1.00	1.00	515
accuracy			0.99	554
macro avg	0.68	0.67	0.68	554
weighted avg	0.99	0.99	0.99	554

Accuracy: 0.9855595667870036

Figure 2: On the left are precision, recall, f1-score, and support for each classification for the SVM model. On the right is the confusion matrix for the SVM model with true labels on the y-axis and predictions on the x-axis.

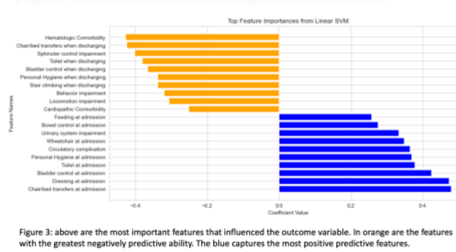


Figure 3: above are the most important features that influenced the outcome variable. In orange are the features with the greatest negatively predictive ability. The blue captures the most positive predictive features.