# Readmission Risk Prediction After Total Hip Arthroplasty Using Machine Learning and Hyperparameter Optimized with Bayesian Optimization

Intan Yuniar Purbasari[1], Athanasius Priharyoto Bayuseno[2], R. Rizal Isnanto[3], Tri Indah Winarni[4]
Doctoral Program of Information System, School of Postgraduate Studies, Diponegoro University, Semarang, Indonesia[1]
Department of Informatics-Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jawa Timur, Surabaya, Indonesia[1]
Department of Mechanical Engineering-Faculty of Engineering, Diponegoro University, Semarang, Indonesia[2]
Department of Computer Engineering-Faculty of Engineering, Diponegoro University, Semarang, Indonesia[3]
Department of Anatomy-Faculty of Medicine, Diponegoro University, Semarang, Indonesia[4]
UNDIP Biomechanics Engineering & Research Centre (UBM-ERC), Universitas Diponegoro, Semarang, Indonesia[4]

***Abstract*—Machine learning techniques are increasingly used in orthopaedic surgery to assess risks such as length of stay, complications, infections, and mortality, offering an alternative to traditional methods. However, model performance varies depending on private institutional data, and optimizing hyperparameters for better predictions remains a challenge. This study incorporates automatic hyperparameter tuning to improve readmission prediction in orthopaedics using a public medical dataset. Bayesian Optimization was applied to optimize hyperparameters for seven machine learning algorithms—Extreme Gradient Boosting, Stochastic Gradient Boosting, Random Forest, Support Vector Machine, Decision Tree, Neural Network, and Elastic-net Penalized Logistic Regression—predicting readmission risk after Total Hip Arthroplasty (THA). Data from the MIMIC-IV database, including 1,153 THA patients, was used. Model performance was evaluated using Precision, Recall, and AUC-ROC, comparing optimized algorithms to those without hyperparameter tuning from previous studies. The optimized Extreme Gradient Boosting algorithm achieved the highest AUC-ROC of 0.996, while other models also showed improved accuracy, precision, and recall. This research successfully developed and validated optimized machine learning models using Bayesian Optimization, enhancing readmission prediction following THA based on patient demographics and preoperative diagnosis. The results demonstrate superior performance compared to prior studies that either lacked hyperparameter optimization or relied on exhaustive search methods.**

***Keywords*—Total hip arthroplasty; orthopaedic surgery; Bayesian Optimization; machine learning algorithm; hyperparameter optimization**

## I. Introduction

Total Hip Arthroplasty (THA) is a surgery to replace the entire hip joint with an artificial joint, called an implant, and is typically performed on patients with severe osteoarthritis and having a prevalent rate of 1.2% [1], [2], [3]. THA is an effective option for reducing pain and improving hip joint function in patients with severe symptomatic osteoarthritis, with a success rate of more than 85% and clinical outcomes lasting for 15 to 25 years [4], [5], [6], [7]. The demand for THA procedures has significantly increased and is expected to continue to rise until 2040 [8], [9], [10]. However, like all surgery procedures, THA comes with risks concerning a patient's condition for infection, dislocation, and mortality and may require unplanned hospital readmission for a corrective procedure or even a revision surgery. Dislocation, complication (both implant and non-implant-associated complication), including infection and mental status, are identified as the reasons for readmission after THA [11], [12]. Meanwhile, 30-day readmission rate of 5% to 9.5% and 5% to 10% for 90-day readmission have also been reported [13], [14], [15], [16], [17], [18].

Bundled payments in healthcare have been established in several countries, such as United States, The Netherlands, Sweden, United Kingdom, and Indonesia, as part of a transition toward value-based treatment. This payment structure makes healthcare providers receive incentives for coordinating treatments, preventing complications and failures, and reducing unnecessary or duplicate tests and treatments, including unplanned hospital readmission [19], [20], [21], [22]. Therefore, a readmission risk prediction system for THA surgery is valuable in not only improving cost management, but also in improving patient care quality. It serves as a tool in the cooperative decision-making process between patients and doctors regarding the final decision to undertake the elective THA surgery, as well as in providing pre- and post-operative treatment tailored to the patient's condition [23], [24], [25]. Based on the predicted risks, in which is very individuals for each patient, doctors may design a customized treatment procedure such as medication prescription, diets, and/or supervised exercise before and after the surgery. The readmission risk prediction system is also a source of information for the hospital to manage the distribution of its resources, such as specialty doctors, operating rooms, inpatient

rooms, and medical equipment [26], [27]. With the rise of medical expenses, implementing effective and efficient procedures across all channels puts hospital management to the test while maintaining the quality of care, as often mirrored by the concept of patient-centered care [28].

Both classical and machine learning-based methods have already applied to predict the risk, including readmission risk, of THA surgery. Classical methods to predict the risk of THA surgery, as well as general surgery, are currently used, such as the Risk Assessment and Predictor Tool (RAPT) [29], Charlson Comorbidity Index (CCI) [30], and Elixhauser Comorbidity Index (ECI) [31], [32]. CCI and ECI measure risks associated with mortality, hospital length of stay, and hospital charges based on the overall severity of comorbidities. Instead, RAPT predicts discharge destination and length of stay for patients after hip or knee arthroplasty. With the rising complexity of patient electronic medical history data, a solution based on artificial intelligence (AI) and machine learning (ML) is critical to understanding each unique health profile of a patient and providing more accurate information related to risks to doctors and patients.

Machine learning (ML) algorithms have been effectively introduced and used in a variety of medical sectors such as predicting length of stay of lung cancer patients [33], elderly patients readmission risk prediction [34], and heart disease classification [35]; nonetheless, its technology is relatively new in the field of orthopaedic surgery [36]. Although classical approaches to calculate complication and mortality risks using patient demographic characteristics and comorbidities do exist, they are considered having weak discrimination, are not verified, and not general enough to accurately predict risk preoperatively and mostly are developed using multivariable regressions techniques [37]. According to the authors' preliminary literature review [38], various ML algorithms have been utilized to predict THA outcomes using AI/ML technology. Among them are Logistic Regression [39], [40], [41], Linear Support Vector Machine [39], [42], Linear Discriminant Analysis [39], Elastic Net Penalized Linear Regression [40], [43], Random Forest [40], [42], [43], Neural Networks [42], [43], Stochastic Gradient Boosting [42], [43], [44], and an automated machine learning tool developed for prognosis (dubbed AutoPrognosis) [45]. Specifically for predicting readmission risk, some machine learning tools used were a Random Forest approach with Natural Language Processing (NLP) feature [46], Linear Regression, Support Vector Machine, Random Forest to predict 90-day readmission [47], Logistic Regression [48], and Light Gradient Boosting Machine [49].

Research by Kuo et al. [50] aimed to predict periprosthetic joint infection by implementing two levels of ML architecture. Naïve-Bayes, eXtreme Gradient Boosting, Linear Regression, and Random Forest were put in the first level as base classifiers, while Support Vector Machine as the meta-classifier acted in the second level as the final decision maker based on the prediction result of the first level. Some patient-related variables used were demographic, biomedical, comorbidity, surgery-related variables and had a very good Area under Curve (AUC) score of 0.988. There was no additional information on hyperparameters fixed for the final classification algorithms in each level.

Research by Klemt [36] investigated the risk of failure of a revision THA involving similar variables to [50] using Neural Network, Elastic-net penalized logistic regression, and Random Forest and achieved the highest AUC score of 0.85 for Neural Network.

The investigation to predict complication and irregular surgery duration was performed in study [44] using eXtreme Gradient Boosting algorithm which incorporated 12 (twelve) variables related to patient demographic, implant type, and surgeon experience. The AUC scores were 0.64 for complication prediction and 0.89 for surgery duration. Based on the supplied codes in GitHub, hyperparameters search was performed using GridSearch method.

Nham et al. [39] conducted a study exploring various tree-based algorithms, neural networks, statistical and probabilistic approaches, and Support Vector Machine (SVM) to predict in-patient mortality, discharge disposition, and length of stay. The study incorporated patient-specific variables, such as demographic, diagnosis, and mortality risk, and also situational variables, such as operation location, hospital type, and number of hospital beds. The best three algorithms were different for each predicted outcome, but Linear Support Vector Machine (LSVM) and Decision List (DL) were the most frequent to appear in the best three. Again, there was no discussion on how to select the best hyperparameters for each algorithm.

A machine learning prediction of all-cause complications within two years of primary THA was introduced by Kunze [42] using preoperative variables, such as demographics, comorbidities, known allergies, and Modified Harris Hip Score. The study investigated several machine learning algorithms such as stochastic gradient boosting, random forest, support vector machine, neural network, and Elastic-net penalized logistic regression, including their hyperparameter settings which were tuned through a grid search approach. The best performing algorithm was Elastic-net penalized logistic regression with an AUC value of 0.93 on test set.

Research by Shah et al. [37] is one of the studies that applied Bayesian Optimization to optimize hyperparameters in an Automated Machine Learning (AutoML) environment. It consists of an ensemble of several machine learning pipelines, each fitted with its own hyperparameters, to make prediction on complications after THA procedure. It achieved AUC value of 0.732 and claimed to exceed the performance of other single-type machine learning algorithms such as Logistic Regression, XGBoost, Gradient Boosting, AdaBoost, and Random Forest.

Specifically for predicting readmission risk, research by Slezak et al. used and compared multiple machine learning techniques such as Light Gradient Boosting Machine (LGBM), Logistic Regression (LR), eXtreme Gradient Boosting (XGB), and Random Forests (RF) to predict risk for readmission after Total Joint Replacement surgery [49]. They included American Society of Anaesthesiologists Physical Status Classification (ASA) and modified Frailty Index (mFI-5) as the contributing

factors and achieved an AUC value of 0.672 with LGBM being the best algorithm for readmission prediction.

Unplanned 90-day readmission after THA was also predicted with LR [48] and the predictive performance of the model was evaluated using the validation dataset, resulting in an AUC value of 0.715. All variables were grouped into five categories, including demographics, perioperative factors, past surgical histories, comorbidities, and medication usages, where contributing factors for each category were identified.

Another prediction of 90-day readmission after TJA (Total Joint Arthroplasty) using comprehensive data from electronic medical records and patient-reported outcome measures was investigated in study [47] by utilizing a variety of machine learning algorithms including LR, LR with LASSO (Least Absolute Shrinkage and Selection Operator) penalty, SVM, and RF. They achieved quite an impressive AUC score of 0.862 with LR-LASSO being the best performing algorithm and selected nine significant variables including diabetes, prior use of corticosteroid medication, and number of follow-up visits to orthopaedic clinics as the significant risk factors.

Most of the machine learning algorithms applied require fine-tuning on a number of the hyperparameters (i.e., the pre-set algorithm parameters to control the learning process and determine the values of model parameters) to achieve the best result (i.e., highest accuracy and precision, lowest error, or highest discriminant value). Finding the best hyperparameters are a manual trial-and-error process, though there has been an increase in using automatic search algorithms (i.e., Hyperparameter Optimization) such as Grid Search, Random Search, Evolutionary-based Search, and Bayesian Optimization. Only a few earlier studies on predicting THA outcomes using machine learning have included automatic hyperparameter tuning with Grid Search [42] and the Bayesian Optimization methodology [37].

In addition, risk prediction of THA patients in earlier studies were based on data from respective authors' affiliations, thus might not be suitable for benchmarking the algorithms' performance and is difficult in terms of reproducibility to verify the results [51]. To the author's knowledge, no previous research related to predicting outcomes following Total Hip Arthroplasty surgery has used public datasets.

Machine learning techniques have proven to be effective in creating models to predict post-operative risks in THA patients, albeit there is no definitive technique best suited for all outcomes. Several previously cited articles briefly discussed how to determine the appropriate hyperparameters for each machine learning algorithm utilized using Grid Search approach, such as [42] and [44], which include systematically testing all possible hyperparameter combinations in the search space, but it is a time consuming and repetitive activity. With the rapidly increasing number and complexity of electronic medical records generated in this big data era, sweeping hyperparameters in the hyperparameter search space thoroughly as Grid Search does would take hours or days as the number of hyperparameters increases and finding the best ones remains a challenge. While Random Search provides better solution by randomly sampling points in the search domain, it is also less efficient for expensive evaluation functions [52]. Another more advanced method such as Bayesian Optimization capable of finding global optimization in a complex problem is gaining more recognition [37], [45]. Consequently, this study aims to provide a Bayesian Optimization approach to MLA in readmission risk predicting after THA operation. This research performed experiments on various machine learning algorithms commonly utilized in existing research with the optimization enhancement on a publicly available medical dataset MIMIC and evaluate the generalizability of the algorithms' claimed performance.

Hyperparameter optimization is particularly advantageous in improving the performance of machine learning algorithms and reducing tedious human effort to find the best setting [53]. It is also widely known that different hyperparameter settings work best for different datasets [54]. In the prediction of THA outcome context, one can expect a more accurate readmission risk prediction while delivering broader and easy-to-use machine learning tools to benefit non-technical expert users (i.e., doctors), incorporating hyperparameter optimization in the machine learning algorithms used. An automated hyperparameter optimization provides the search for the best hyperparameter configuration for each machine learning algorithm used automatically, thus improving accuracy and efficiency, as well as lowering error [55]. Meanwhile, automation makes it easier and simpler for non-technical experts to use various machine-learning technologies and gain more from the outcomes [53], [56]. The suggested system's contributions include the following:

*1) Development of readmission risk prediction method*: The study proposes to incorporate automatic hyperparameters tuning to the machine learning algorithm selected by users, providing a more accurate and precise prediction result while offering ease and simplicity in applying and customizing machine learning algorithms to non-technical expert users.

*2) Exploration of hyperparameters optimization algorithm based on statistical approach*: The study introduces a Bayesian Optimization method which is based on statistical calculation to find the best hyperparameters setting of the machine learning algorithm selected by users.

*3) Performance evaluation on known public tabular dataset (Medical Information Mart for Intensive Care-IV) MIMIC-IV*: To evaluate the generalizability of high performing machine learning algorithms acclaimed in previous studies, this study uses a known public medical dataset MIMIC-IV [57], which is based on real medical dataset.

*4) Performance evaluation with robust metrics*: to assess the effectiveness of the proposed system, the study introduces four evaluation metrics: Accuracy, Precision, Recall, and Area under the Curve Receiver Operating Characteristic (AUC-ROC) values. These metrics measure the algorithm's performance quality quantitatively.

This journal article is organized as follows: Section I introduces the background, objectives, and significance of the study; Section II details the methods, including data sources,

preprocessing, and machine learning techniques; Section III presents the results of the experiments; Section IV discusses the findings in comparison to existing research; and Section V concludes with key takeaways and future research directions.

## II. METHODS

### A. System Overview

The proposed readmission risk prediction system after THA operation furnishes decision support feature to both patients and healthcare professionals in the domain of peril prognosis. The system processes medical data from the electronic medical record dataset which integrates various relevant attributes from a number of tables. All related data from patients underwent THA procedure are extracted, including demographics, diagnosis, and postoperative procedures of medical rehabilitation, if any. Additionally, some data imputations of missing values from the attributes were performed and the readmission status of patients within a year after the THA procedure is set to be the target value of prediction. The system includes a varied array of classification algorithms in previous studies, including eXtreme Gradient Boosting (XGB), Stochastic Gradient Boosting (SGB), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Neural Network (NN), and Elastic-net Penalized Logistic Regression (EnPLR). The best hyperparameters of those algorithms are then selected using Bayesian Optimization (BO) technique to ensure maximum results of the classification algorithms in predicting the readmission risk. Evaluation metrics such as Precision, Recall, and Area under the Receiver Operating Characteristic (AUROC) curve are employed to measure the algorithms' performance quality. Fig. 1 explicates the system overview of the proposed readmission risk prediction system.

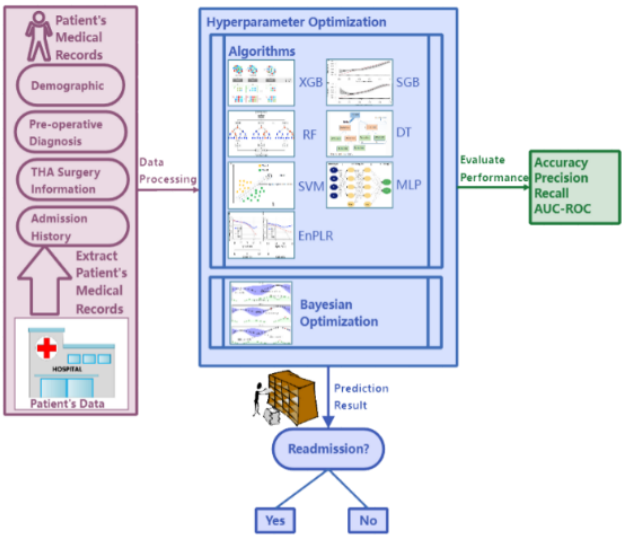



Fig. 1. System overview of optimized readmission risk prediction model.

The proposed method predicts the readmission risk of a THA patient based on one's demographics and preoperative diagnosis to help doctors plan preoperative treatments, aid patients in making informed decision on the surgery, and to assist hospitals managing their resources efficiently and properly with data-driven decision making. The system involved steps as follows:

*1)* Preprocess the patient dataset from MIMIC-IV (Medical Information Mart for Intensive Care version IV, which is a large de-identified public dataset of patients admitted to the emergency unit at the Beth Israel Deaconess Medical Center in Boston, MA). Duplicate data is identified and removed.

*2)* Identify the hyperparameters of previously utilized machine learning algorithms, including XGB, SGB, RF, SVM, DL, NN, and EnPLR, and perform the Bayesian Optimization technique to find the best hyperparameters for each algorithm.

*3)* Divide the dataset into 70% training and 30% testing set with stratified sampling.

*4)* Compare the model performance using evaluation metrics such as Precision, Recall, and AUROC.

### B. Algorithm Introduction

*1) eXtreme Gradient Boosting (XGB)*: XGB is a tree boosting machine learning algorithm whose capability of handling sparse data becomes one of its highlight features. The algorithm was introduced by Chen and Guestrin [58] and has won various machine learning competitions. It implements gradient boosting technique to perform additive optimization and incorporates a regularized model to prevent overfitting using L1 and L2 regularization methods. The algorithm is available in various popular languages such as Python, R, C, C++, and Java.

*2) Stochastic Gradient Boosting (SGB)*: Introduced long before XGB, SGB has different approach to prevent overfitting in a general gradient boosting technique. It proposed a randomness into the algorithm by drawing a subsample, which is a fraction $f$ of the size of the training set, to replace the full training data set at each iteration [59]. In scikit-learn, SGB is implemented with GradientBoostingClassifier by defining subsample hyperparameter < 1.

*3) Random Forest (RF)*: RF works by building multiple decision trees and combines the output to give a single result. RF technique extends the bagging approach by combining bagging and feature randomness to generate an uncorrelated forest of decision trees [60]. Some of key benefits of RF are its capability to handle classification and regression tasks, and it is also easy to determine feature importance from the prediction result. This advantage can help to identify which features contribute the most to the prediction made.

*4) Support Vector Machine (SVM)*: SVM has the ability to perform both linear and non-linear classification using a kernel trick. It works by constructing a hyperplane in a high dimension space and determine functional margin of data points (called support vectors) which define a separation boundary between classes [61].

*5) Decision Tree (DT)*: In this study, DT is used as a replacement of Decision List (DL), since a Python's scikit-learn implementation of DL is not available. DT is a simpler

version of RF which consists of decision nodes, chance nodes, and end nodes. Following the path from root and making decisions at every branch according to the attribute value leading to an end node will create a rule that defines the temporal or causal relations among attributes.

*6) Neural network*: The Neural Network implemented in this study is Multi-Layer Perceptron (MLP), which is a supervised learning algorithm that maps a number of inputs to a number of outputs with a number of nodes in between (called hidden layers). The nodes in the hidden layers transform the values from the previous layer by adjusting the nodes' weights and propagate the result to the next layer. MLP implements a regularization technique (called L2) to avoid overfitting by punishing weights of large magnitudes.

*7) Elastic-net Penalized Logistic Regression (EnPLR)*: EnPLR is a modified version of the classic Logistic Regression with a regularization method to overcome overfitting that linearly combines L1 (LASSO-Least Absolute Shrinkage and Selection Operator) and L2 (Ridge Regression). In scikit-learn, EnPLR is implemented with LogisticRegression model by setting hyperparameters solver=saga, penalty=elasticnet, and l1_ratio=0.5.

*8) Hyperparameter optimization with Bayesian Optimization (BO)*: Hyperparameter optimization is a process to fine tuning the hyperparameter values of machine learning algorithms to obtain the best setting which yield the best result, in a classification or regression problem. The domains of hyperparameter may range from real values (i.e. learning rate), integer (i.e. number of layers), binary (i.e. with or without early stopping), to category (i.e. a selection of optimizer functions). Let $\mathcal{A}$ be a machine learning algorithm with N hyperparameters. The domain of $n$-th hyperparameter is denoted as $\Lambda_n$ and the overall configuration hyperparameter space is denoted as $\Lambda = \Lambda_1 \, x \, \Lambda_2 \, x \, ... \, \Lambda_N$. A hyperparameter vector is denoted as $\lambda \in \Lambda$ and a machine learning algorithm with its instantiated hyperparameter $\lambda$ is denoted as $\mathcal{A}_\lambda$. Given a dataset $\mathcal{D}$, the goal of hyperparameter optimization is to find Eq. (1) as in [53]:

$$\lambda^* = \underset{\lambda \in \mathcal{A}}{argmin} \, \mathbb{E}_{(D_{train}, D_{valid}) \sim \mathcal{D}} \boldsymbol{V}(\mathcal{L}, \mathcal{A}_\lambda, D_{train}, D_{valid}) \quad (1)$$

with $\boldsymbol{V}(\mathcal{L}, \mathcal{A}_\lambda, D_{train}, D_{valid})$ measures the loss value of the model generated by algorithm A with hyperparameter λ on the training data $D_{train}$ and evaluated on validation data $D_{valid}$. $D \sim \mathcal{D}$ denoted the finite data which its expectation needs to be approximated.

Grid Search and Random Search are the two most basic and simple hyperparameter optimization methods. While Grid Search performs exhaustive search on the hyperparameter search space, Random Search [56] samples a hyperparameter configuration randomly until it satisfies a defined threshold value. On the other hand, Bayesian Optimization (BO) is a global optimization framework approach of a black box objective function with high complexity and undefined shape and offers better efficiency given limited evaluation budget [62], [63].

BO is used in hyperparameter optimization (HO) that constructs a probability model of an objective function and use it to select the most promising hyperparameters to be evaluated in the real objective function [64]. Utilizing a Bayesian approach, BO takes note of the result of the previous evaluation to build a probability model which maps hyperparameters onto a probability score of an objective function, *P(score|hyperparameters)*, called a surrogate function. Compared to the original function, the new function is easier to be evaluated and the Bayesian method works by finding the next hyperparameter having the best performance on the surrogate function to be evaluated on the real objective function. In a BO, a Gaussian Process (GP) is used as a surrogate function to produce a posterior distribution over the function values based on observed data.

Meanwhile, an acquisition function selects the next point for the surrogate function, with common options including probability of improvement, expected improvement, and upper confidence bounds. This study uses the Bayesian Optimization library *bayes-opt*, which defaults to upper confidence bounds to balance exploitation of high surrogate values and exploration of uncertain regions.

### C. Datasets

The proposed study utilized public medical dataset from MIMIC-IV (Medical Information Mart for Intensive Care version IV) [62] from Beth Israel Deaconess Medical Centre in Boston, MA, United States. The dataset consists of 223,452 patients in its hospitalization hosp module and after a careful selection resulted in 1,153 unique THA patients' data. The target variable is readmission within one year of first THA procedure which is a Boolean value TRUE or FALSE.

A comprehensive description of characteristics of the patient data is provided in Table I. Most of the demographic-related variables are summarized as mean values, while the rest are written based on their respective numbers and percentage. The numbers in preoperative diagnosis variable do not sum up to the total number of patients as a patient may have more than one diagnosis.

### D. Preprocessing

Data preprocessing involves identifying and removing duplicates, missing values, and erroneous information from the data. Proper techniques such as Label Encoding and One-Hot Encoding are applied to transform categorical data into a numerical format [63]. Additionally, a standard scaler is applied to ensure all attribute values are standardized. The overall preprocessing step is summarized in Fig. 2.

Data preprocessing involved extracting relevant data from the general dataset, identifying 1,153 unique THA patients with 55 variables (54 inputs and 1 output), as summarized in Table I. Patients had up to 39 preoperative diagnoses, totalling 1,881 distinct diagnoses coded in International Classification of Diseases ICD-9 and ICD-10.

In addition, a feature processing was applied. Missing values in weight, height, and Body Mass Index (BMI) were handled by filling with the mean value of the attribute column. All 1,881 diagnosis ICDs were assigned integer values from 1 to 1881 (a value of 9999 was given to empty cells), implant

materials variable in categorical type were assigned 1 to 7 (99 for empty cells), implant type in categorical were assigned values of 1 to 4 (9 for empty cells), and insurance type in categorical were assigned values of 1 to 3.

TABLE I. CHARACTERISTICS OF THA PATIENTS

| Variables | All Patients (n=1,153) | |
|---|---|---|
| | Mean | (IQR) |
| *Demographics* | | |
| Age (years) | 65.66 | 16 (58-74) |
| Weight (pounds) | 180.91 | 65.75 (146.25-212) |
| Height (feet) | 43.67 | 67 (0-67) |
| Body Mass Index (kg/m$^2$) | 28.98 | 8.05 (24.6-32.65) |
| *Preoperative Blood pressure* | | |
| Systole | 131.62 | 20 (120-140) |
| Diastole | 74.88 | 15 (67-82) |
| | **Number** | **(%)** |
| Male Sex | 506 | 43.88 |
| *Insurance type* | | |
| Medicare | 505 | 43.80 |
| Medicaid | 35 | 3.03 |
| Other | 613 | 53.16 |
| *Preoperative diagnosis* | | |
| Osteoarthritis | 1,023 | 8.9 |
| Hypertension | 579 | 5.04 |
| Hyperlipidemia | 440 | 3.83 |
| Esophageal reflux | 336 | 3.83 |
| Nicotine/tobacco dependence | 207 | 2.93 |
| Anemia | 338 | 2.94 |
| Depressive disorder | 111 | 0.97 |
| Anxiety disorder | 154 | 1.34 |
| Sleep apnea | 200 | 1.74 |
| Diabetes | 260 | 2.26 |
| Asthma | 155 | 1.35 |
| Others | 7,679 | 66.88 |
| *Implant material* | | |
| Synthetic | 196 | 16.99 |
| Metal on Polyethylene | 85 | 7.37 |
| Ceramic on Polyethylene | 437 | 37.90 |
| Ceramic | 31 | 2.69 |
| Ceramic on Ceramic | 4 | 0.34 |
| Metal | 6 | 0.52 |
| Oxidized Zirconium on Polyethylene | 2 | 0.17 |
| N/A | 392 | 33.99 |
| *Implant type* | | |
| Cemented | 36 | 3.12 |
| Cementless | 300 | 26.02 |
| Not Specified | 217 | 18.82 |
| N/A | 600 | 52.04 |
| *Readmission within one-year post-THA* | 237 | 20.55 |

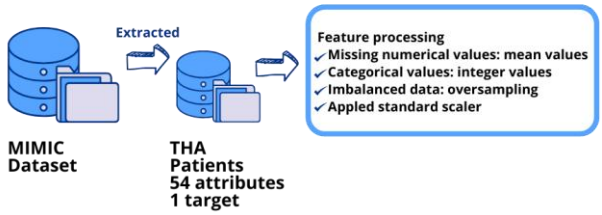



Fig. 2. Preprocessing steps.

Since the dataset is imbalanced, with 237 TRUE and 916 FALSE, a simple random oversampling method to duplicate examples in the minority class was applied using *imbalanced-learn* library [64]. The oversampling method produced equal size of data for each target class, which were 916 TRUE and 916 FALSE.

## III. RESULTS

### A. Algorithm Selection and Hyperparameter Optimization (HO)

The experiment was conducted in two modes for each algorithm: without and with hyperparameter optimization (HO). A limitation of the bayes-opt library is its restriction to numeric attributes, preventing the optimization of categorical and Boolean hyperparameters. Five numerical hyperparameters were optimized for each algorithm, except for Support Vector Machine and Elastic-net Penalized Logistic Regression, which had only three due to categorical constraints. Table II presents the optimized hyperparameters, their values, and a comparison of accuracy and AUROC between Bayesian Optimization (BO) and default hyperparameters.

The convergence plot in Fig. 3 shows that six of seven algorithms (except SVM) reached a steady value within six iterations, indicating BO has effectively estimated the optimization target. Even SVM's sharp increase in the 14th iteration [Fig. 3(d)] remained within the 95% Confidence Interval (CI).

The objective plot in Fig. 4 visualizes the distribution of hyperparameter values when the BO searched for the optimized ones. The diagonal subplots represent histograms of the sampled values for each parameter and aid in visualizing the distribution of parameter values used during optimization. This illustrates how BO investigated the parameter space. The red dashed line represents the optimized hyperparameter value while the blue line plots the AUC-ROC value as the objective target around the optimized hyperparameter value. If the majority of a parameter's samples are concentrated in a specific range, it may indicate that this range has values that are near ideal. For example, four straight lines in the diagonal subplots of XGB [Fig. 4(a)] indicates that all the sampled values of expGamma (an exponential transformation of Gamma), learning_rate, max_depth, and min_child_weight between their respective ranges yield a near optimal AUC-ROC value, i.e. 0.9955. Meanwhile, the off-diagonal contour subplots show pairwise interactions between two hyperparameters, where darker colors indicate better values. The contour identifies regions where the parameter combinations yield optimal results. For example, the off-diagonal subplot between expGamma and expC (the exponential transformation of C) in Fig. 4(c) indicates that the red star value ($expGamma = 0.584$, thus $Gamma = 10^{0.584} = 3.837$) in the dark region represents the optimal value of expGamma when combined with the value of expC. In Fig. 4(a), the red star in n_estimators plot (n_estimators=1818.9744) resides within the narrow darker area, indicating that the value is optimal or near optimal.

TABLE II. THE HYPERPARAMETERS OF EACH ALGORITHM AND THEIR OPTIMIZED VALUES

| Algorithm | Hyperparameters | Bounds | | optimized value | without HO | | with HO | |
|---|---|---|---|---|---|---|---|---|
| | | min | max | | auc-roc | fit time (s) | auc-roc | elapsed time (s) |
| XGB | learning_rate | 0.01 | 0.8 | 0.0352 | 0.9955 | 0.18 | 0.9958 | 582.11 |
| | min_child_weight | 0 | 5 | 1.5814 | | | | |
| | max_depth | 1 | 50 | 50 | | | | |
| | gamma | 1e-5 | 1.0 | 0.0731 | | | | |
| | n_estimators | 5 | 5000 | 1818.9744 | | | | |
| | | | | | | | | |
| SGB | learning_rate | 0.01 | 0.8 | 0.4091 | 0.8790 | 0.72 | 0.9936 | 747.03 |
| | n_estimators | 10 | 250 | 151.6208 | | | | |
| | subsample | 0.1 | 0.9 | 0.8513 | | | | |
| | min_sample_split | 2 | 25 | 11.3614 | | | | |
| | max_depth | 0 | 500 | 155.4370 | | | | |
| | | | | | | | | |
| RF | max_depth | 0 | 500 | 34.0807 | 0.9945 | 0.66 | 0.9936 | 541.20 |
| | max_features | 0.1 | 0.999 | 0.3961 | | | | |
| | max_leaf_nodes | 0 | 5000 | 1556.2286 | | | | |
| | min_samples_split | 2 | 25 | 6.0441 | | | | |
| | n_estimators | 10 | 250 | 219.7180 | | | | |
| | | | | | | | | |
| SVM | C | 1e-6 | 1e+6 | 1230.18891 | 0.7655 | 0.25 | 0.9915 | 6060.32 |
| | gamma | 1e-4 | 1e+5 | 2.359238 | | | | |
| | tol | 1e-9 | 0.1 | 0.1 | | | | |
| | | | | | | | | |
| DT | max_depth | 1 | 500 | 179.5508 | 0.8782 | 0.03 | 0.8783 | 4.75 |
| | min_samples_split | 2 | 25 | 10.5158 | | | | |
| | min_samples_leaf | 1 | 5 | 3.8508 | | | | |
| | max_features | 0.1 | 0.999 | 0.5504 | | | | |
| | max_leaf_nodes | 2 | 1000 | 715.1166 | | | | |
| | | | | | | | | |
| MLP | alpha | 1e-6 | 10 | 0.0226 | 0.9142 | 3.12 | 0.9503 | 485.52 |
| | learning_rate_init | 1e-6 | 10 | 0.0012 | | | | |
| | batch_size | 10 | 300 | 65.5406 | | | | |
| | max_iter | 100 | 1000 | 801.9782 | | | | |
| | tol | 1e-9 | 0.01 | 0.0003 | | | | |
| | | | | | | | | |
| EnPLR | tol | 1e-9 | 0.01 | 0.0000002 | 0.6592 | 1.67 | 0.6622 | 923.93 |
| | C | 0.001 | 1000 | 526.6542 | | | | |
| | max_iter | 5000 | 10000 | 8923.7024 | | | | |

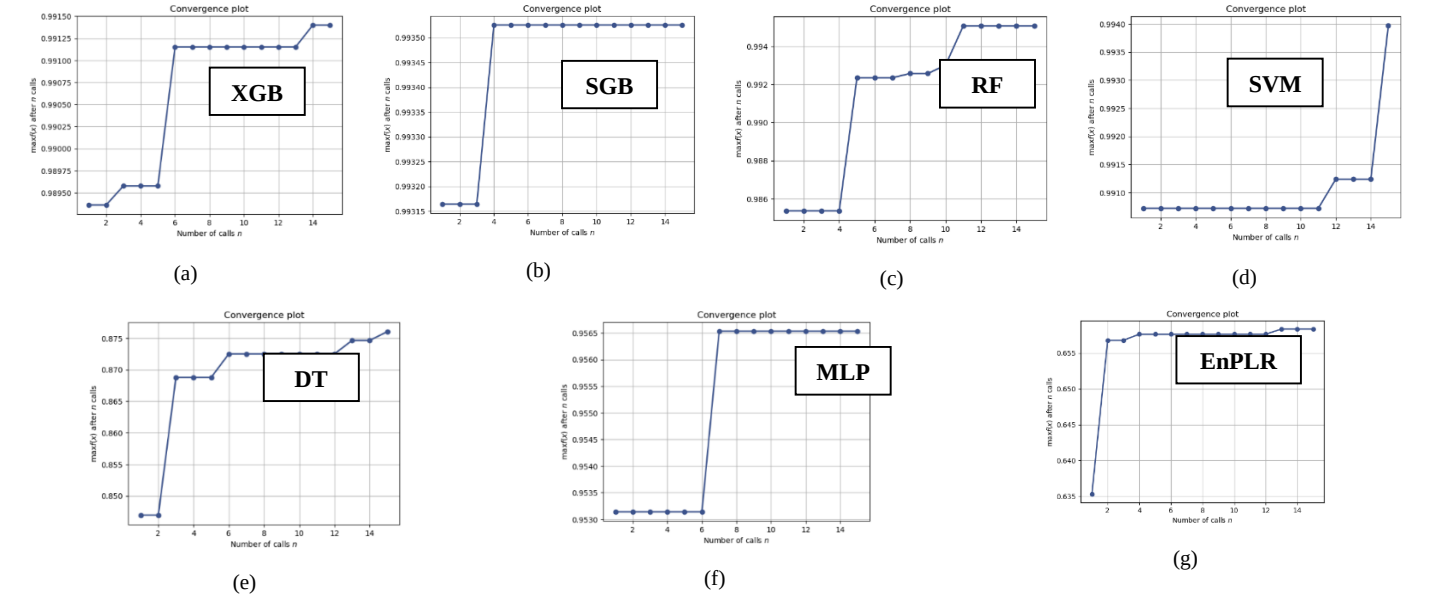
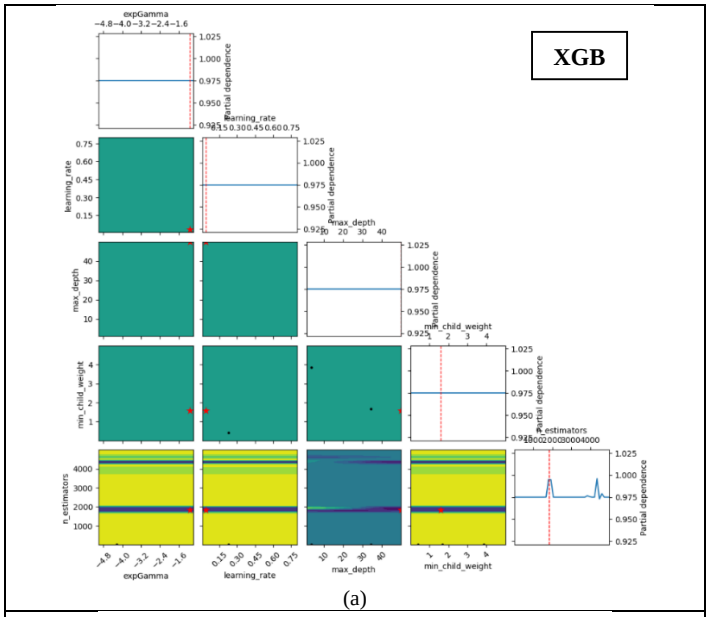
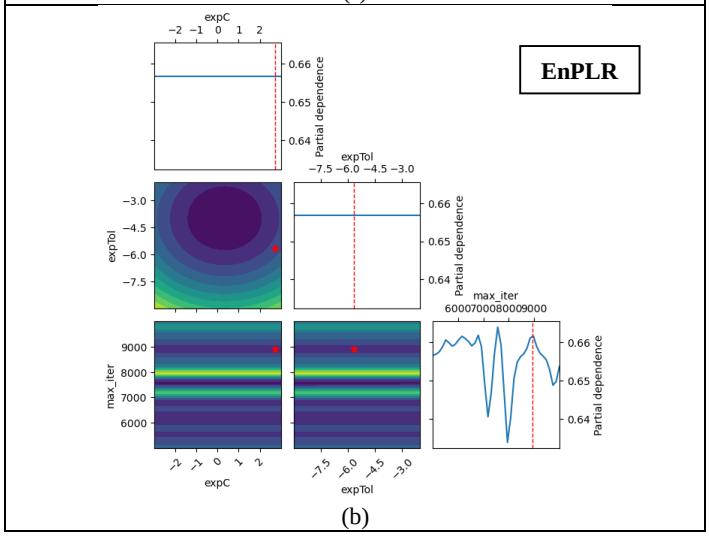


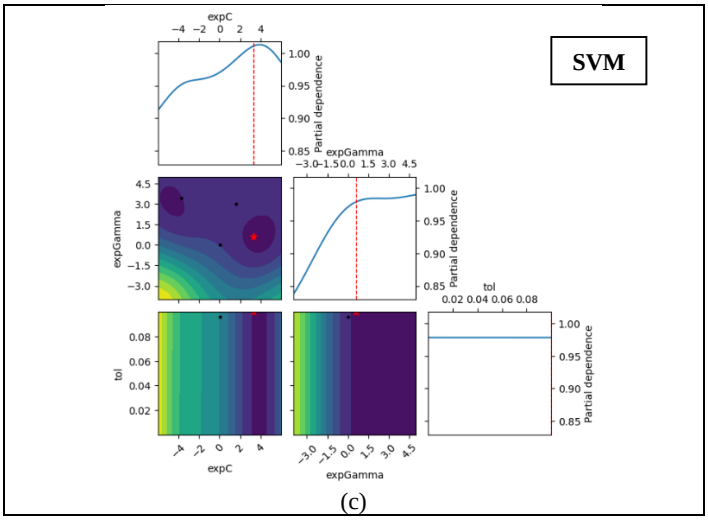

Fig. 3. The convergence plot in bayesian optimization technique for algorithm: (a) Extreme gradient boosting (XGB), (b) Stochastic gradient boosting (SGB), (c) Random forest (RF), (d) Support vector machine (SVM), (e) Decision tree (DT), (f) Multilayer perceptron (MLP), and (g) Elastic-net penalized linear regression (EnPLR).




(a)




(b)




(c)

Fig. 4. The objective plot of (a) XGB, (b) EnPLR, and (c) SVM. The red stars represent the best values of each hyperparameter.

## B. *Performance Evaluation*

Table II shows the time required for each approach to complete optimization. All seven algorithms took longer with Bayesian Optimization (BO) than with default hyperparameters, with execution times increasing from 155 times longer (MLP) to 24,605 times longer (SVM). This added duration is expected in hyperparameter tuning, balancing computational cost with improved AUC-ROC performance.

Table III presents performance metrics: Accuracy, Precision, Recall, and AUC-ROC. SGB, SVM, and MLP showed improvements in all four metrics after optimization, while RF and DT performed worse. Optimization had minimal impact on XGB and EnPLR.

Additionally, MLP's default setting (max_iter = 200) failed to converge, requiring an increase to max_iter = 1000 in the optimized version. Similarly, EnPLR failed to converge in both cases, even after raising max_iter to 10,000 in the optimized version.

TABLE III. PERFORMANCE EVALUATION COMPARISON OF ALGORITHM

| Alg. | Accuracy | | Precision | | Recall | | AUROC | |
|---|---|---|---|---|---|---|---|---|
| | w/o HO | with HO | w/o HO | with HO | w/o HO | with HO | w/o HO | with HO |
| XGB | 0.945 | 0.950 | 0.908 | 0.916 | 0.991 | **0.991** | 0.995 | **0.996** |
| SGB↑ | 0.801 | 0.976 | 0.767 | **0.971** | 0.866 | 0.982 | 0.879 | 0.994 |
| RF↓ | 0.978 | 0.967 | 0.965 | 0.947 | 0.991 | 0.989 | 0.994 | 0.994 |
| SVM↑ | 0.693 | **0.991** | 0674 | 0.947 | 0.750 | 0.989 | 0.764 | 0.994 |
| DT↓ | 0.878 | 0.820 | 0.814 | 0.780 | 0.985 | 0.894 | 0.878 | 0.878 |
| MLP↑ | 0.862 | 0.900 | 0.813 | 0.843 | 0.942 | 0.986 | 0.914 | 0.950 |
| EnPLR | 0.630 | 0.632 | 0.623 | 0.624 | 0.668 | 0.672 | 0.660 | 0.662 |

↑: an increased value when using BO; ↓: a decreased value when using BO

## IV. Discussion

Previous THA-related studies have employed machine learning techniques using institution-specific datasets, contributing to research advancements. However, many relied on Grid Search for hyperparameter tuning or default algorithm settings, limiting optimization efficiency. Only a few studies explored advanced methods such as Bayesian Optimization (BO) to enhance predictive performance, despite its potential for complex optimization problems.

This study applies BO across multiple machine learning algorithms commonly used in previous research to improve readmission prediction. Our results demonstrate that optimized hyperparameters significantly enhance AUC-ROC scores compared to default settings. While prior studies addressed different adverse event predictions, they shared a common challenge: imbalanced datasets, as adverse events are rare.

Comparing model performance, this study outperformed previous research across several algorithms. The Extreme Gradient Boosting (XGB) algorithm achieved the highest AUC-ROC score (0.996), surpassing [44], other cited works [36], [37], [39], [42], [50], and specifically [49] for predicting readmission risk. Similarly, the Multilayer Perceptron (MLP) achieved an AUC-ROC of 0.914, higher than the Neural Network in [36] (0.85). The Support Vector Machine (SVM) in this study achieved 0.994, far outperforming previous studies [39], which reported 0.74 (Length of Stay), 0.8 (Discharge), and 0.97 (Mortality). The Stochastic Gradient Boosting (SGB) algorithm improved to 0.994, higher than 0.88 in [42]. Random Forest (RF) achieved 0.994, surpassing 0.80 in [36], 0.91 in [42], and 0.83 in [47] for readmission prediction. However, Elastic-Net Penalized Logistic Regression (EnPLR) underperformed (0.662), falling below 0.732 in [37], 0.93 in [42], and 0.86 in [47]. These findings highlight the effectiveness of Bayesian Optimization in tuning machine learning models, addressing limitations in previous studies that relied on basic or exhaustive hyperparameter searches.

Additionally, the MIMIC-IV dataset used in this study provides a publicly available benchmark to validate models trained on private institutional datasets, which often restrict benchmarking and generalizability. SVM, which performed best in some studies [39], [50], showed weaker performance before optimization, while XGB, previously considered suboptimal [39], [44], demonstrated superior results. This suggests that models trained on private datasets may not always generalize well to broader patient populations.

Furthermore, Table III indicates that Recall scores are consistently higher than Accuracy and Precision, suggesting that models prioritize identifying positive cases, reducing false negatives. This is critical in medical applications, where minimizing false negatives (missed readmission cases) is more important than false positives, which can be addressed with further evaluation [65], [66], [67].

The discrepancy in comparative results between datasets may stem from differences in patient populations, clinical settings, and data preprocessing. The MIMIC-IV dataset includes a more diverse patient cohort from multiple institutions, whereas private datasets from previous studies may be more homogeneous. Additionally, the inclusion/exclusion criteria and feature selection may introduce variability in model performance.

The varying performance of models across datasets suggests that some algorithms are more sensitive to dataset structure. For instance, tree-based models (XGB, RF, SGB) performed consistently well, likely due to their robustness in handling structured tabular data. SVM, which performed best in some previous studies, showed inferior performance before optimization in our dataset, suggesting that hyperparameter tuning plays a crucial role in improving its adaptability to different data distributions. Our results show that Bayesian Optimization significantly improves performance consistency across datasets. Without proper hyperparameter tuning, models such as SVM and MLP may underperform in certain datasets due to suboptimal parameter selection. This highlights the importance of adaptive optimization techniques in ensuring machine learning models generalize well across different clinical datasets.

Despite its contributions, this study has several limitations. First, as a retrospective analysis relying on past medical records and billed ICD codes, it is susceptible to errors such as miscoding or missing values, which could impact machine learning predictions. Second, the bayes-opt library used for hyperparameter tuning is still under active development, and its latest version (1.5.1 as of July 10, 2024) does not yet support categorical and Boolean hyperparameters. As a result, key parameters such as SVM kernels, MLP activation functions, and RF criteria could not be optimized, potentially limiting model performance. Lastly, the dataset exhibited class imbalance, with significantly fewer TRUE cases than FALSE cases. While this study employed oversampling to address the issue, more advanced techniques could further enhance predictive accuracy. Future research should explore alternative resampling strategies and hyperparameter tuning methods to refine model performance.

## V. Conclusion

This study successfully developed, trained, tested, and validated Bayesian Optimization for hyperparameter tuning across seven machine learning algorithms to predict readmission risk following Total Hip Arthroplasty (THA). By comparing these optimized models to previous studies that either lacked hyperparameter tuning or employed different optimization methods, the results demonstrated that Bayesian Optimization significantly enhances predictive performance. This underscores the critical role of hyperparameter tuning in maximizing model accuracy, improving decision-making, and increasing confidence in integrating machine learning into THA patient management.

Additionally, this study highlights the importance of external evaluation using a publicly available dataset, ensuring that machine learning models trained on institution-specific data can generalize effectively across different patient populations. The findings suggest that standardized evaluation is essential for ensuring model robustness and reliability in clinical applications.

While this study addressed key challenges in predictive modeling for THA readmission, limitations remain, particularly regarding class imbalance and the scope of readmission prediction. Future research may explore advanced resampling techniques to further improve model performance and investigate more specific predictive outcomes, such as infection-related readmission, revision THA, or Length of Stay.

## ACKNOWLEDGMENT

This research is supported by School of Postgraduate Studies-Diponegoro University and University of Pembangunan Nasional Veteran Jawa Timur.

## ETHICAL APPROVAL

This study used the MIMIC-IV database (Medical Information Mart for Intensive Care, version 2.0), a publicly available and de-identified dataset, which complies with the Health Insurance Portability and Accountability Act (HIPAA) standards. Access to the dataset was granted after completing the Collaborative Institutional Training Initiative (CITI) Program training and acceptance of the Data Use Agreement (DUA). Ethical approval for the original data collection was obtained by the Institutional Review Board (IRB) of Beth Israel Deaconess Medical Center (BIDMC), and the requirement for individual patient consent was waived. Therefore, no additional ethical approval was required for this study.

## REFERENCES

[1] Australian Orthopaedic Association, "Australian Orthopaedic Association National Joint Replacement Registry," 2022. [Online]. Available: https://aoanjrr.sahmri.com/annual-reports-2022].

[2] H.-H. Bleß Miriam Kip Eds, "White Paper on Joint Replacement," Berlin, 2018. doi: https://doi.org/10.1007/978-3-662-55918-5.

[3] R. Brittain et al., "NJR statistical analysis, support and associated services," 2021. [Online]. Available: www.njrcentre.org.uk

[4] J. T. Evans, J. P. Evans, R. W. Walker, A. W. Blom, M. R. Whitehouse, and A. Sayers, "How long does a hip replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up," 2019. [Online]. Available: www.thelancet.com

[5] X. Y. Mei, Y. J. Gong, O. Safir, A. Gross, and P. Kuzyk, "Long-term outcomes of total hip arthroplasty in patients younger than 55 years: A systematic review of the contemporary literature," 2019, Canadian Medical Association. doi: 10.1503/cjs.013118.

[6] I. P. Sitorus and I. H. Dilogo, "Functional and Radiological Outcome of Revision Total Hip Arthroplasty in the Indonesian National Referral Hospital," Hip Knee J, vol. 3, no. 1, pp. 2723–7826, 2022, doi: 10.46355/hipknee.v3i1.161.

[7] M. Spalević et al., "TOTAL HIP REPLACEMENT REHABILITATION: RESULTS AND DILEMMAS," Acta Medica Medianae, vol. 57, no. 1, pp. 48–53, Mar. 2018, doi: 10.5633/amm.2018.0108.

[8] S. M. Kurtz, K. L. Ong, E. Lau, and K. J. Bozic, "Impact of the Economic Downturn on Total Joint Replacement Demand in the United States: Updated Projections to 2021," JBJS, vol. 96, no. 8, 2014, [Online]. Available: https://journals.lww.com/jbjsjournal/Fulltext/2014/04160/Impact_of_the_Economic_Downturn_on_Total_Joint.2.aspx

[9] J. A. Singh, S. Yu, L. Chen, and J. D. Cleveland, "Rates of Total Joint Replacement in the United States: Future Projections to 2020–2040 Using the National Inpatient Sample," J Rheumatol, vol. 46, no. 9, p. 1134, Sep. 2019, doi: 10.3899/jrheum.170990.

[10] M. Sloan, A. Premkumar, and N. P. Sheth, "Projected Volume of Primary Total Joint Arthroplasty in the U.S., 2014 to 2030," JBJS, vol. 100, no. 17, 2018, [Online]. Available: https://journals.lww.com/jbjsjournal/fulltext/2018/09050/projected_volume_of_primary_total_joint.3.aspx

[11] M. A. Smolle et al., "Readmissions at Thirty-Days and One-Year for Implant-Associated Complications following Primary Total Hip and Knee Arthroplasty: A Population-Based Study of 34,392 Patients Across Austria," J Arthroplasty, Aug. 2024, doi: 10.1016/j.arth.2024.08.027.

[12] R. C. Clement et al., "Risk factors, causes, and the economic implications of unplanned readmissions following total hip arthroplasty," Journal of Arthroplasty, vol. 28, no. 8 SUPPL, pp. 7–10, 2013, doi: 10.1016/j.arth.2013.04.055.

[13] R. M. Greiwe, J. M. Spanyer, J. R. Nolan, R. N. Rodgers, M. A. Hill, and R. G. Harm, "Improving Orthopedic Patient Outcomes: A Model to Predict 30-Day and 90-Day Readmission Rates Following Total Joint Arthroplasty," J Arthroplasty, vol. 34, no. 11, pp. 2544–2548, Nov. 2019, doi: 10.1016/j.arth.2019.05.051.

[14] S. Zhao, J. Kendall, A. J. Johnson, A. A. G. Sampson, and R. Kagan, "Disagreement in Readmission Rates After Total Hip and Knee Arthroplasty Across Data Sets," Arthroplast Today, vol. 9, pp. 73–77, Jun. 2021, doi: 10.1016/j.artd.2021.04.002.

[15] D. E. Goltz et al., "A Novel Risk Calculator Predicts 90-Day Readmission Following Total Joint Arthroplasty," JBJS, vol. 101, no. 6, 2019, [Online]. Available: https://journals.lww.com/jbjsjournal/fulltext/2019/03200/a_novel_risk_calculator_predicts_90_day.9.aspx

[16] S. M. Kurtz, E. C. Lau, K. L. Ong, E. M. Adler, F. R. Kolisek, and M. T. Manley, "Has Health Care Reform Legislation Reduced the Economic Burden of Hospital Readmissions Following Primary Total Joint Arthroplasty?," J Arthroplasty, vol. 32, no. 11, pp. 3274–3285, 2017, doi: https://doi.org/10.1016/j.arth.2017.05.059.

[17] J. Williams, B. S. Kester, J. A. Bosco, J. D. Slover, R. Iorio, and R. Schwarzkopf, "The Association Between Hospital Length of Stay and 90-Day Readmission Risk Within a Total Joint Arthroplasty Bundled Payment Initiative," J Arthroplasty, vol. 32, no. 3, pp. 714–718, 2017, doi: https://doi.org/10.1016/j.arth.2016.09.005.

[18] M. A. Varacallo, L. Herzog, N. Toossi, and N. A. Johanson, "Ten-Year Trends and Independent Risk Factors for Unplanned Readmission Following Elective Total Joint Arthroplasty at a Large Urban Academic Hospital," J Arthroplasty, vol. 32, no. 6, pp. 1739–1746, 2017, doi: https://doi.org/10.1016/j.arth.2016.12.035.

[19] J. N. Struijs, E. F. De Vries, C. A. Baan, P. F. Van Gils, and M. B. Rosenthal, "Bundled-Payment Models Around the World: How They Work and What Their Impact Has Been," 2020. Accessed: Oct. 26, 2024. [Online]. Available: https://www.commonwealthfund.org/sites/default/files/2020-04/Struijs_bundled_payment_models_around_world_ib.pdf

[20] R. E. Mechanic, "Mandatory Medicare Bundled Payment — Is It Ready for Prime Time?," New England Journal of Medicine, vol. 373, no. 14, pp. 1291–1293, Oct. 2015, doi: 10.1056/nejmp1509155.

[21] C. Bazell, M. Alston, P. M. Pelizzari, and B. A. Sweatman, "BUndled payment US," pp. 1–5, Mar. 27, 2023. Accessed: Oct. 26, 2024. [Online]. Available: https://www.milliman.com/en/insight/what-are-bundled-payments-and-how-can-they-be-used-by-healthcare-organizations

[22] MInister of Health of the Republic of Indonesia, Regulation of the Minister of Health of the Republic of Indonesia concerning Guidelines for the Implementation of the National Health Insurance Program. Jakarta: MInister of Health of the Republic of Indonesia, 2014.

[23] M. A. Fontana, S. Lyman, G. K. Sarker, D. E. Padgett, and C. H. MacLean, "Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty?," Clin Orthop Relat Res, vol. 477, no. 6, pp. 1267–1279, 2019, doi: 10.1097/CORR.0000000000000687.

[24] M. Huber, C. Kurz, and R. Leidl, "Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning," BMC Med Inform Decis Mak, vol. 19, no. 1, p. 3, 2019, doi: 10.1186/s12911-018-0731-6.

[25] J. Sniderman, R. B. Stark, C. E. Schwartz, H. Imam, J. A. Finkelstein, and M. T. Nousiainen, "Patient Factors That Matter in Predicting Hip Arthroplasty Outcomes: A Machine-Learning Approach," Journal of Arthroplasty, vol. 36, no. 6, pp. 2024–2032, 2021, doi: 10.1016/j.arth.2020.12.038.

[26] HealthCare.gov, "Payment Bundling."

[27] NEJM Catalyst, "What Are Bundled Payments?," Catalyst Carryover, vol. 4, no. 1, Aug. 2023, doi: 10.1056/CAT.18.0247.

[28] Institute of Medicine (US) Committee on Quality of Health Care in America, Crossing the Quality Chasm: A New Health System for the 21st Century. Washington (DC; US): National Academies Press, 2001. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25057539

[29] L. B. Oldmeadow, H. McBurney, and V. J. Robertson, "Predicting risk of extended inpatient rehabilitation after hip or knee arthroplasty," J. Arthroplasty, vol. 18, no. 6, pp. 775–779, 2003, doi: 10.1016/s0883-5403(03)00151-7.

[30] M. E. Charlson, P. Pompei, K. L. Ales, and C. R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation," J Chronic Dis, vol. 40, no. 5, pp. 373–383, 1987, doi: https://doi.org/10.1016/0021-9681(87)90171-8.

[31] A. Elixhauser, C. Steiner, D. R. Harris, and R. M. Coffey, "Comorbidity measures for use with administrative data," Med. Care, vol. 36, no. 1, pp. 8–27, 1998, doi: 10.1097/00005650-199801000-00004.

[32] D. E. Goltz, S. P. Ryan, C. B. Howell, D. Attarian, M. P. Bolognesi, and T. M. Seyler, "A Weighted Index of Elixhauser Comorbidities for Predicting 90-day Readmission After Total Joint Arthroplasty," Journal of Arthroplasty, vol. 34, no. 5, pp. 857–864, May 2019, doi: 10.1016/j.arth.2019.01.044.

[33] B. Alsinglawi et al., "An explainable machine learning framework for lung cancer hospital length of stay prediction," Sci Rep, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-021-04608-7.

[34] S. D. Mohanty, D. Lekan, T. P. McCoy, M. Jenkins, and P. Manda, "Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare," Patterns, vol. 3, no. 1, Jan. 2022, doi: 10.1016/j.patter.2021.100395.

[35] R. R. Isnanto, I. Rashad, and C. Edi Widodo, "Classification of Heart Disease Using Linear Discriminant Analysis Algorithm," in E3S Web of Conferences, R. R. Isnanto, H. null, and B. Warsito, Eds., EDP Sciences, 2023. doi: 10.1051/e3sconf/202344802053.

[36] C. Klemt et al., "Can machine learning models predict failure of revision total hip arthroplasty?," Arch Orthop Trauma Surg, Jun. 2022, doi: 10.1007/s00402-022-04453-x.

[37] A. A. Shah, S. K. Devana, C. Lee, R. Kianian, M. van der Schaar, and N. F. SooHoo, "Development of a Novel, Potentially Universal Machine Learning Algorithm for Prediction of Complications After Total Hip Arthroplasty," Journal of Arthroplasty, vol. 36, no. 5, pp. 1655-1662.e1, 2021, doi: 10.1016/j.arth.2020.12.040.

[38] I. Yuniar Purbasari, A. Priharyoto Bayuseno, R. R. Isnanto, T. Indah Winarni, and J. Jamari, "Artificial Intelligence and Machine Learning in Prediction of Total Hip Arthroplasty Outcome: A Bibliographic Review," in E3S Web of Conferences, R. R. Isnanto, H. null, and B. Warsito, Eds., EDP Sciences, 2023. doi: 10.1051/e3sconf/202344802054.

[39] F. H. Nham, T. Court, A. K. Zalikha, M. M. El-Othmani, and R. P. Shah, "Assessing the predictive capacity of machine learning models using patient-specific variables in determining in-hospital outcomes after THA," J Orthop, vol. 41, pp. 39–46, 2023, doi: 10.1016/j.jor.2023.05.012.

[40] C. Klemt et al., "The utility of machine learning algorithms for the prediction of patient-reported outcome measures following primary hip and knee total joint arthroplasty," Arch Orthop Trauma Surg, vol. 143, no. 4, pp. 2235–2245, 2023, doi: 10.1007/s00402-022-04526-x.

[41] O. Pakarinen, M. Karsikas, A. Reito, O. Lainiala, P. Neuvonen, and A. Eskelinen, "Prediction model for an early revision for dislocation after primary total hip arthroplasty," PLoS One, vol. 17, no. 9 September, 2022, doi: 10.1371/journal.pone.0274384.

[42] K. N. Kunze, A. V Karhade, E. M. Polce, J. H. Schwab, and B. R. Levine, "Development and internal validation of machine learning algorithms for predicting complications after primary total hip arthroplasty," Arch Orthop Trauma Surg, vol. 143, no. 4, pp. 2181–2188, 2023, doi: 10.1007/s00402-022-04452-y.

[43] K. N. Kunze, A. V. Karhade, A. J. Sadauskas, J. H. Schwab, and B. R. Levine, "Development of Machine Learning Algorithms to Predict Clinically Meaningful Improvement for the Patient-Reported Health State After Total Hip Arthroplasty," Journal of Arthroplasty, vol. 35, no. 8, pp. 2119–2123, Aug. 2020, doi: 10.1016/j.arth.2020.03.019.

[44] I. Lazic et al., "Prediction of Complications and Surgery Duration in Primary Total Hip Arthroplasty Using Machine Learning: The Necessity of Modified Algorithms and Specific Data," J Clin Med, vol. 11, no. 8, 2022, doi: 10.3390/jcm11082147.

[45] A. M. Alaa and M. van der Schaar, "AutoPrognosis: Automated Clinical Prognostic Modeling via Bayesian Optimization with Structured Kernel Learning," in Proceedings of the 35th International Conference on Machine Learning, J. Dy and A. Krause, Eds., Stockholm, Sweden, Jul. 2018, pp. 139–148.

[46] V. Digumarthi et al., "Preoperative prediction model for risk of readmission after total joint replacement surgery: a random forest approach leveraging NLP and unfairness mitigation for improved patient care and cost-effectiveness," J Orthop Surg Res, vol. 19, no. 1, Dec. 2024, doi: 10.1186/s13018-024-04774-0.

[47] J. Park et al., "Machine Learning-Based Predictive Models for 90-Day Readmission of Total Joint Arthroplasty Using Comprehensive Electronic Health Records and Patient-Reported Outcome Measures," Arthroplast Today, vol. 25, Feb. 2024, doi: 10.1016/j.artd.2023.101308.

[48] M. Korvink, C. W. Hung, P. K. Wong, J. Martin, and M. J. Halawi, "Development of a Novel Prospective Model to Predict Unplanned 90-Day Readmissions After Total Hip Arthroplasty," Journal of Arthroplasty, vol. 38, no. 1, pp. 124–128, Jan. 2023, doi: 10.1016/j.arth.2022.07.017.

[49] J. Slezak, L. Butler, and O. Akbilgic, "The role of frailty index in predicting readmission risk following total joint replacement using light gradient boosting machines," Inform Med Unlocked, vol. 25, Jan. 2021, doi: 10.1016/j.imu.2021.100657.

[50] F. C. Kuo, W. H. Hu, and Y. J. Hu, "Periprosthetic Joint Infection Prediction via Machine Learning: Comprehensible Personalized Decision Support for Diagnosis," Journal of Arthroplasty, vol. 37, no. 1, pp. 132–141, Jan. 2022, doi: 10.1016/j.arth.2021.09.005.

[51] H. Shaheen, P. Marimuthu, and D. Rashmi, "The Practical Approaches of Datasets in Machine Learning," in AI Applications in Cyber Security and Communication Networks, C. Hewage, L. Nawaf, and N. Kesswani, Eds., Singapore: Springer Nature Singapore, 2024, pp. 221–227.

[52] Y. Rimal, N. Sharma, and A. Alsadoon, "The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms," Multimed Tools Appl, vol. 83, no. 30, pp. 74349–74364, 2024, doi: 10.1007/s11042-024-18426-2.

[53] M. Feurer and F. Hutter, "Hyperparameter Optimization," in Automated Machine Learning: Methods, Systems, Challenges, F. Hutter, L. Kotthoff, and J. Vanschoren, Eds., Cham: Springer International Publishing, 2019, pp. 3–33. doi: 10.1007/978-3-030-05318-5_1.

[54] R. Kohavi and G. H. John, "Automatic Parameter Selection by Minimizing Estimated Error," in Machine Learning Proceedings 1995, A. Prieditis and S. Russell, Eds., San Francisco (CA): Morgan Kaufmann, 1995, pp. 304–312. doi: https://doi.org/10.1016/B978-1-55860-377-6.50045-1.

[55] A. Goodfellow, Ian; Bengio, Yoshua; Courville, Deep Learning. The MIT Press, 2016.

[56] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," Journal of Machine Learning Research, vol. 13, no. 10, pp. 281–305, 2012, [Online]. Available: http://jmlr.org/papers/v13/bergstra12a.html

[57] A. E. W. Johnson et al., "MIMIC-IV, a freely accessible electronic health record dataset," Sci Data, vol. 10, no. 1, p. 1, 2023, doi: 10.1038/s41597-022-01899-x.

[58] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. ACM, Aug. 2016. doi: 10.1145/2939672.2939785.

[59] J. H. Friedman, "Stochastic gradient boosting," Comput Stat Data Anal, vol. 38, no. 4, pp. 367–378, 2002, doi: https://doi.org/10.1016/S0167-9473(01)00065-2.

[60] L. Breiman, "Random Forests," Mach Learn, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[61] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[62] A. Johnson et al., "MIMIC-IV (version 3.0)," 2024, PhysioNet. doi: https://physionet.org/content/mimiciv/3.0/.

[63] D. Chicco, "Ten quick tips for machine learning in computational biology," BioData Min, vol. 10, p. 35, 2017, doi: 10.1186/s13040-017-0155-3.

[64] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning," Journal of Machine Learning Research, vol. 18, no. 17, pp. 1–5, 2017, [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[65] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," Artif Intell Med, vol. 23, no. 1, pp. 89–109, 2001, doi: https://doi.org/10.1016/S0933-3657(01)00077-X.

[66] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in Proceedings of the 23rd International Conference on Machine Learning, in ICML '06. New York, NY, USA: Association for Computing Machinery, 2006, pp. 233–240. doi: 10.1145/1143844.1143874.

[67] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Brief Bioinform, vol. 19, no. 6, pp. 1236–1246, 2018, doi: 10.1093/bib/bbx044.