Data preprocessing and applied machine learning on rheumatology study

Seyedsiamak Rouzmeh
Friedrich-Alexander-Universität
Erlnagen-Nürenberg
siamak.rouzmeh@fau.de

Abstract

In this study the use of machine learning (ML) techniques for predicting the effectiveness of methotrexate (MTX) treatment on rheumatoid arthritis (RA) patients is presented. The objective is to create a predictive algorithm using clinical, serologic, genomic and sociodemographicat data at early RA patient to predict the effectiveness of drug in 24 weeks before continuing the treatment with other synthetic or biologic antirheumatic drugs such as bdMARDs or tsDMARDs. The study focuses on using iterative imputation instead of simple imputation of missing value with mean or median as well as developing the model performance and computation speed through using optuna method for hyperparameter tuning. Moreover, we were able to apply few popular and practical supervised machine learning models including Randomforest(RF), Adaboost, XGboost, SVM and lightGBM to find out which model performs best for our case and also extract the most important features which influencing the patients responds to treatment via SHAP value summary result. Due to limited study subjects(330 patients in 6 month Therapy period) and to avoid data leakage we implemented nested cross validation via hyperparamter tunning in inner loop and feature selection and model evaluation in outer loop. All model evaluation matrices showed that RF and XGboost models are outperforming other candidates in predicting the correct responders and non-responders to MTX treatment. Besides charachteristics such as DAS28 value, HAQScore, ESR, RF, physician activity and ages are determining most the patients responds.

1. Introduction

Rheumatoid Arthritis (RA) is a disease which pain in joints is a highlighted symptom. Many researches have been done to reduce the treatment cost, optimizing the drug consumption and identifying effective clinical properties on remission rate [13].

On the other hand, applied Machine Learning(ML) has been widely used in medical applications in recent decays including Medical imaging analysis, drug discovery and development, personalized therapy and etc. More specifically, ML enables the development of personalized treatment plans based on individual patient characteristics. By analyzing patient data, including genetic information, medical history, and lifestyle factors, machine learning models can provide tailored treatment recommendations to optimizing the outcomes and minimizing side effects [3, 6].

ML models are categorized to supervised and unsupervised model which train labeled and non-labeled data respectively. In case to predict the therapy responds we need to have the proper label for the data with enouough records of patients. The more data available from patients, the better ML will perform to support the rheumatologist. Generally, there are two main challenges in manipulating the data, first is the quantity and second is the quality of the data. Rheumatologists will likely benefit from information from national cohorts, local registries, or larger datasets from electronic medical records (EMRs). The data quality could be effected due to noisy data, missing values, and irregular visits which lower the model's overall quality. To solve this issue, data preprocessing including, data cleaning and filling missing value should be done to provide high quality data [3, 15].

Currently, MTX (combination) therapy is applied to all patients with early RA for at least 3-6 months, allowing time for the build-up dose and a reliable assessment of the drug's effectiveness. Patients whom do not respond to the MTX will continue their therapy with targeted synthetic or biologic antirheumatic drugs (ts/bDMARDs) [11]. Delays of three to six months can cause patients to lose the window of opportunity for successful treatment of RA disease activity and expose themselves to potential MTX-related side effects. Therefore, finding baseline predictors to show which patients are likely to respond well to MTX and stay on MTX used as monotherapy would be essential. The motivation to use Machine Learning models as a tool to predict these important personalizes characteristic based on the available data is a big and essential step towards an individualised approach to RA treatment [11].

A study early RA patient(n¿5000) treated by MTX showed that ML methods Comparing manual modelling to integrating baseline clinical data did not significantly enhance the prediction of MTX treatment persistence at 12 months and the highest area under the curve (AUC) for least absolute shrinkage and selection operator (LASSO) regression was 0.67 [16]. Another study with 335 patients reveals that In patients with RA who took MTX as monotherapy or in combination with other DMARDs, the performance of ML algorithms (LASSO models AUC 0.76) was not superior to the logistic regression (AUC 0.77) in predicting DAS28 more than 3.2 at 3 months of treatment [5]. Duong *et al*. [4] studied on 775 patients respond to MTX basen on DAS28-ESR value criteria for remission. Features such as DAS28-ESR, , anti-citrullinated protein antibody (ACPA), and Health Assessment Questionnaire (HAQ) were recognized as top features for responders. In the external validation set, they obtained results through training with LASSO (area under the curve [AUC] 0.79] and random forests [AUC 0.68]).

In this study, we aimed to increase the model performance to predict the respondanc and non-respondace in 6 months tratment. Since we initially investigated on increasing the data quality through choosing suitable imputation method and then we developed a malgorithm that the feature selection and hyperparameter tunning can be implemented in nested cross validation. Subsequently, we tried different supervised ML models to get the best fit to extract important baseline characteristic effecting the MTX treatment procedure.

2. Materials and Methods

2.1. Data acquisition and participants

The follow-up data of 387 patients treated with methotrexate was gathered from Erlangen-University Hospital.

2.2. Clinical outcomes and definition of response

Sociodemographic, clinical, and genomic data at baseline were used to predict response to methotrexate treatment. Patients who took methotrexate after in 6-month ptherapy and experienced either a good or moderate response based on the EULAR response criteria are referred to be responders to methotrexate monotherapy. The response criteria were based on the effectiveness of the treatment which is DAS28-ESR score less than 3.2.

2.2.1 Pre-processing

The common approach to fill the missing value is to use the simple imputer which fill the values based on the average, median or other simple statistical value of that specific feature. This method is not accurate estimation for missing data. Another candidate is using the iterative imputation(MICE) to uses other features value to predict the missing value data via linear regression fitting. This more complicated approach provide more accurate estimation which at the end increase the model performance. In our case, we tested several methods to handle missing data on the data set and chose the one that produced the best outcomes based on our evaluation criteria. We applied chained equations (MICE), k-nearest-neighbors-based imputation (KNN), mean and median imputation [10]. Among those MICE method is more accurate in filling missing value, specially because the algorithm replaces missing values with estimated values based on the other observed data.

After obtaining high quality data-set we did labeling based on DAS28 value for patients who has at-least 6 month follow-up period. Patients with DAS28 value equal or less than 3.2 value get positive effectiveness label as assumed to be responders.

To balance the number of subject for each label we applied resampling method via SMOTE function which make the responders and non-responders subject equivalent via oversampling.

2.3. Model design

To avoid overfitting, data leakage, and to use all training data, nested cross-validation(CV) with 5 iteration was used to train the classifiers by tuning model hyperparameters in inner CV and feature selection in outer CV. According to figure 1 Feature selection was done in outer loop to increase the accuracy and also to determine important Sociodemographic and genomic factors which influence the prediction result. For this task we used embedded feature selection in each predictor model, except, for model which not embedded with feature selection, the random forest feature selection was used.

The hyperparamter tunning in inner loop is also shown in figure 1. A popular candidate for hyperparameter tunning is determining the set of values for each paramter and search through Grid-search. Grid search explore the search space exhaustively and explore all possible combinations of these values within the defined search space. However, this process can be computationally expensive and not suitable for algorithm with many hyperparamters. A A robust alternative is tunning hyperparamter by optuna method via creating an objective function and define trial. Optuna uses Bayesian optimisation, a more sophisticated and adaptive strategy. Optuna iteratively explores the search space using a method known as sequential model-based optimisation (SMBO) based on the knowledge obtained from prior iterations. Optuna chooses a selection of hyperparameter configurations to examine based on their potential to enhance

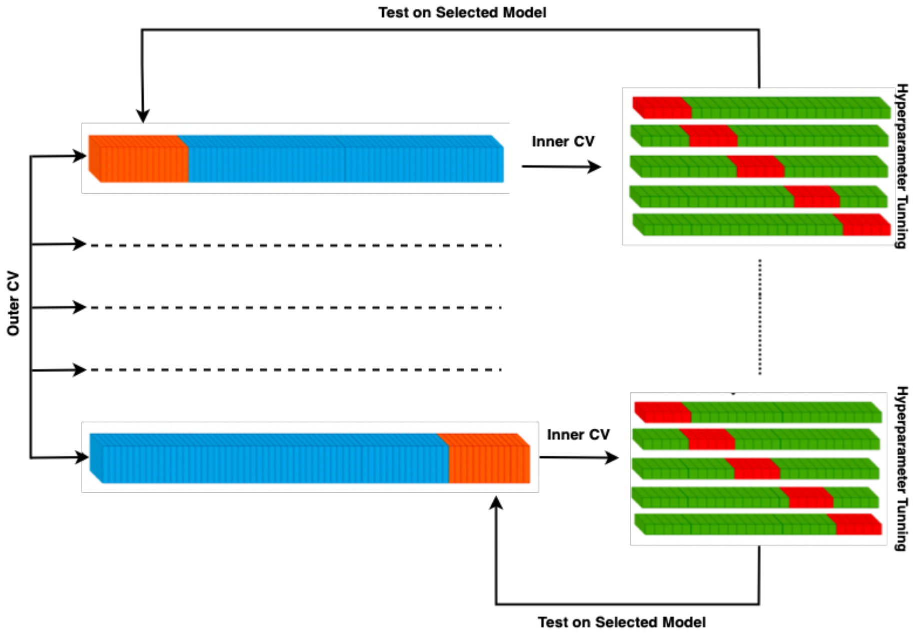


Figure 1: Machine Learning strategy

the performance of the model rather than evaluating every possible combination. this method focuses its search efforts on areas of the search space that are most likely to produce better results by utilising statistical models and Bayesian inference. It can locate the ideal parameter configuration more quickly and with fewer evaluations than grid search due to this adaptive technique [1, 2]. In our study, we defined objective function with different search space for each model which maximizes the mean accuracy calculated in each inner fold. The number of trial was set to 50 for all models.

After finding the best parameters for the selected features, five supervised ML classification methods including Adaboost, XGboost, Randomforest, Lightgbm and SVM were implemented to identify predictors of good response to MTX treatment. To guarantee the results' replication, only the best trained model was assessed on the external validation dataset.

2.3.1 Performance Evaluation

The performance was measured by calculating different matrices such as accuracy, area under curve (AUC) of a receiver operating characteristic curve (ROC), precision-recall curve, F1 score in outer CV of the nested cross validation.

2.3.2 Feature importance analysis

For the predictions' interoperability, the Shapely additive explanation (SHAP) value was applied.

An influence on a patient's response to treatment is shown by a positive (or negative) Shapely value. Shapely values that are higher signify stronger influences on patient response, and vice versa.

3. Result

3.1. Baseline sociodemographics and clinical factors

The filtered dataset contain 330 subjects which has 43 features. Among them 197 are responders and 133 are non-responders. The demographic and baseline clinical characteristics are summarized in Tables 1.

3.2. Model Performances

The evaluation matrices including average accuracy, F1-score and precision of all models are plotted in figure 2. Among all matrices only the averaged recall value show significantly better performance of XGboost and Random forest, while other matrices show non-comparable performance. Therefore we provide the box plot for accuracy and f1-score in figure 2 (a) and (b) respectively to have a better comparison of model performance. It is clear from the box plots that the Random Forest and XGBoost classifiers perform significantly better than others. Comparing Random

Table 1: Demographics and baseline clinical characteristics of the RA patients

BaselineCharacteristics	Overall(n = 330)	Responder(n=197)	Non_responder(n=133)
Age_years	55.92(13.44)	53.24(13.57)	59.89(12.25)
Gender(Female,%)	233(71%)	115(57%)	118(89%)
BMI	28.59(17.14)	27.76(13.79)	29.82(21.14)
Hypertension(%)	85(26%)	40(20%)	45(34%)
Depression(%)	19(6%)	9(5%)	10(8%)
Diabetes(%)	22(7%)	17(9%)	5(4%)
Fat_metabolism_disorder(%)	21(6%)	12(6%)	9(7%)
Gout(%)	12(4%)	8(4%)	4(3%)
Osteoporosis(%)	18(5%)	4(2%)	14(11%)
Thyroid_disease(%)	60(18%)	37(19%)	23(17%)
SJC28	1.67(2.44)	1.44(1.91)	2.01(3.03)
TJC28	2.62(3.28)	2.33(3.2)	3.05(3.35)
VAS_activity_physician	29.56(45.33)	31.07(45.48)	27.32(45.17)
VAS_activity_patient	32.16(13.52)	31.77(13.61)	32.74(13.41)
VAS_pain	30.98(28.58)	29.27(35.38)	33.52(12.91)
HAQ_Score	0.98(0.9)	0.88(1.09)	1.14(0.46)
DAS28ESR_Score	3.23(0.73)	3.1(0.65)	3.42(0.8)
DAS28CRP_Score	3.11(0.67)	3.07(0.6)	3.16(0.76)
CDAI_Score	10.46(4.19)	10.05(2.44)	11.07(5.86)
SDAI_Score	12.29(4.69)	11.68(3.29)	13.21(6.11)
ESR_mm	22.71(15.1)	20.9(18.36)	25.39(7.49)
CRP_mg_l	0.89(0.31)	0.88(0.33)	0.91(0.29)
RF	139.13(216.94)	108.06(150.67)	185.16(282.88)
CCP	0.54(8.97)	0.88(11.61)	0.05(0.23)
tsDMARD(%)	2(1%)	1(1%)	1(1%)
bDMARD(%)	39(12%)	23(12%)	16(12%)

Data are shown in mean (standard deviation) if not otherwise specified

Forest and XGBoost to the other models, the clear distinction between the median lines and the interquartile ranges shows that these models have higher and more consistent prediction accuracy [8, 9].

Table 2: Performance metrics of different classifiers

Classifier	Accuracy	F1 score	Recall	Precision
Adaboost	0.751	0.746	0.751	0.765
RF	0.761	0.758	0.842	0.771
SVM	0.726	0.722	0.778	0.735
XGboost	0.767	0.761	0.839	0.778
lightgbm	0.754	0.748	0.819	0.766

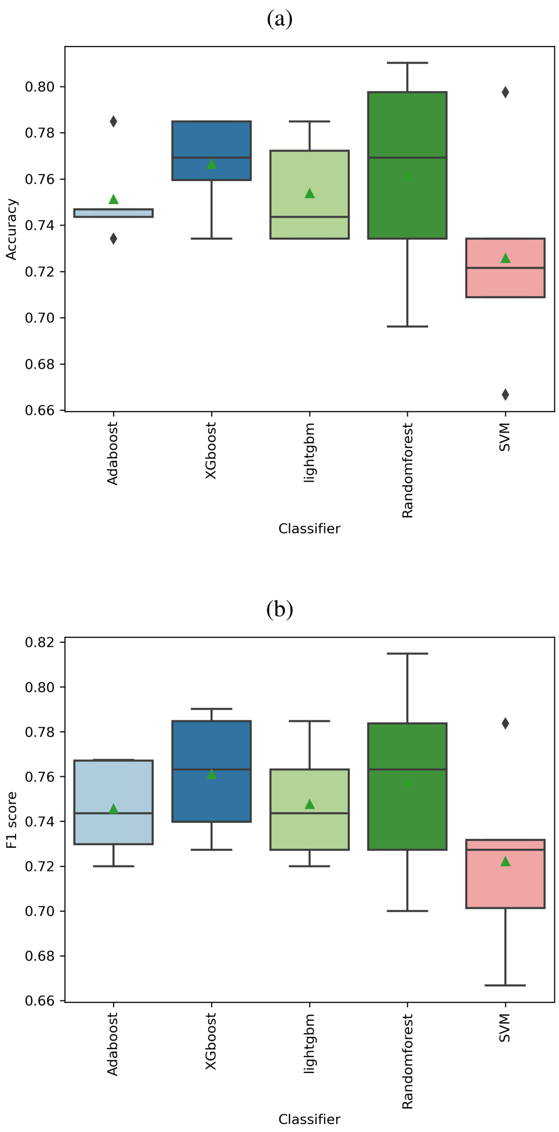
AUC-ROC curves tell us how the model can distinguish between different classes in prediction. The higher area under curve indicates the better classification performance of the model. From figure 3, we can notice that Random forest and XGboost classifier both with AUC 0.84 are performing better than other models in classification. Lightgbm shows slightly lower AUC value(0.82) which leads to it's fair performance in distinguishing the right classes. On the other hand, SVM with AUC 0.78 and Adaboost with 0.75 signifficatly show lower performance due to the lower AUC-ROC value.

After figuring out best two models for our case study, we plot the calibration curve as it is shown in figure 4. A potential discrepancy between the probability predicted by the model and the probabilities observed in the data is displayed in such plot. The plot curve is closer to this straight line for better calibration. Obviously, random forest is better calibrated compared to XGboost as it is closer to the straight line [14].

Confusion Matrix for XGboost model is available in figure 5. The predicted and actual class designations are broken down in detail by a confusion matrix. This matrix can be used as a suitable representative way of the model performance instead of Precision, recall or F1-score as they are derived from information in Confusion [12].

3.3. Characteristics importance

The SHAP value summary plot for XGboost classifier is shown in figure 6.The SHAP value illustrates the con-


(a)

(b)

Figure 2: (a) Accuracy and (b) F1-score box plot for all models

tribution of each feature in model prediction. The model's predictions are more strongly influenced by features with higher SHAP values than by those with lower SHAP values. Also, the sign of SHAP value shows whether it has positive or negative contribution. The vertical bar indicate that how the magnitude of each feature can impact in negative or positive way [7].

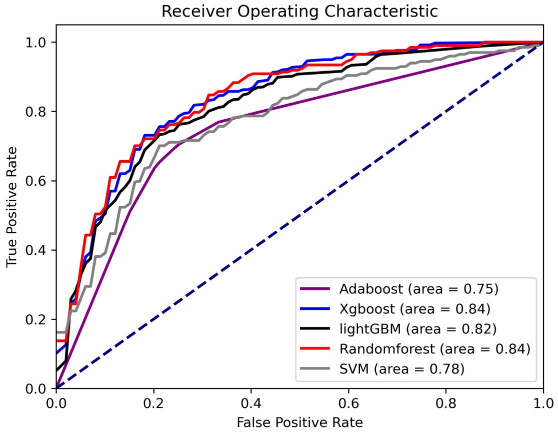


Figure 3: ROC curve for all classification models

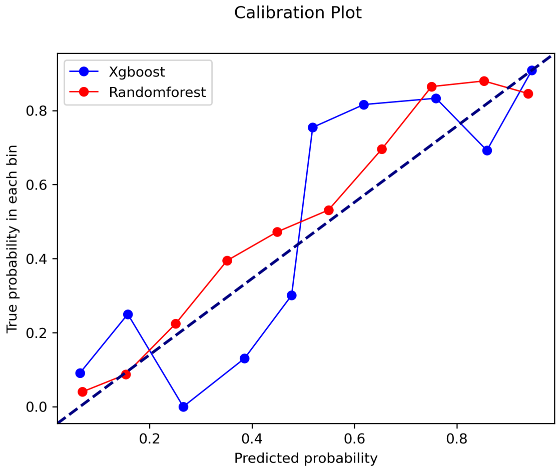


Figure 4: The calibration curve for Xgboost and Randomforest classifiers

4. Discussion

Based on model performance matrices, specially AUC-ROC value, XGboost and Randomforest outperform other models, particularly ADaboost and SVM. The reason could be related to the fact that These two models can capture complex relationship between features and they are more robust to the noises and outliers. Moreover, the calibration plot in figure 4 shows that the hyperparmeter tuning step was done properly in these models specially RF as it is closed to the straight line.

Comparing to other studies [16, 4], we coudld achieved AUC-ROC 0.84 for both XGboost and RF which shows an improvement in predicting the respond of the RA patient to MTX therapy.

Figure 6 shows that features such as DAS28-ESR, HAQ, ESR, RF, VAS-Activity-Physician, Age, Das28CRP, SDAI, VAS-Activity-patient, Gender, TJC28, BMI and VAS-pain could be important predictors that influence the effectiv-

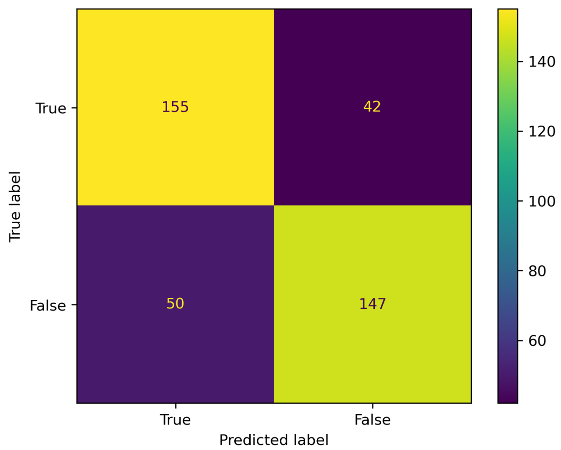


Figure 5: The confusion matrix for XGboost classifier

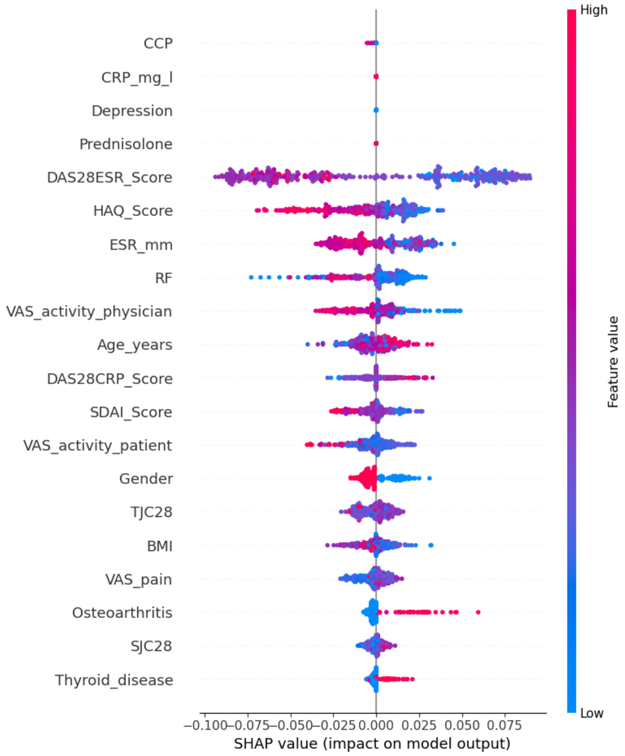


Figure 6: SHAP value plot for XGboost classifier

nesss of the MTX therapy. There are meaningfull and worthy interpretation that worth mentioning from SHAP summary plot.

5. Conclusion

As summary and conclusion, we tested and developed different supervised Machine Learning models on Patients with early RA disease which were under treatment for 6 month visit to predict the effectiveness of the MTX drug in this period. The Random forest and XGboost classifier perform better in classification task rather than other models.

References

[1] T. Bex. Why is everyone at kaggle obsessed with optuna for hyperparameter tuning?, Aug 2021.

[2] T. Bex. You are missing out on lightgbm. it crushes xgboost in every aspect, Sep 2021.

[3] G. R. Burmester. Rheumatology 4.0: big data, wearables and diagnosis by computer. *Annals of the rheumatic diseases*, 77(7):963–965, 2018.

[4] S. Q. Duong, C. S. Crowson, A. Athreya, E. J. Atkinson, J. M. Davis, K. J. Warrington, E. L. Matteson, R. Weinshilboum, L. Wang, and E. Myasoedova. Clinical predictors of response to methotrexate in patients with rheumatoid arthritis: a machine learning approach using clinical trial data. *Arthritis Research & Therapy*, 24(1):1–11, 2022.

[5] H. R. Gosselt, M. M. Verhoeven, M. Bulatović-Ćalasan, P. M. Welsing, M. C. de Rotte, J. M. Hazes, F. P. Lafeber, M. Hoogendoorn, and R. de Jonge. Complex machine-learning algorithms and multivariable logistic regression on par in the prediction of insufficient clinical response to methotrexate in rheumatoid arthritis. *Journal of personalized medicine*, 11(1):44, 2021.

[6] M. Hügle, P. Omoumi, J. M. van Laar, J. Boedecker, and T. Hügle. Applied machine learning and artificial intelligence in rheumatology. *Rheumatology advances in practice*, 4(1):rkaa005, 2020.

[7] D. Kirk. Using shap with cross-validation in python, Dec 2022.

[8] Z. LT. Essential things you need to know about f1-score, Nov 2021.

[9] Z. LT. Precision and recall made simple, Nov 2021.

[10] M. Mera-Gaona, U. Neumann, R. Vargas-Canas, and D. M. López. Evaluating the impact of multivariate imputation by mice in feature selection. *Plos one*, 16(7):e0254720, 2021.

[11] E. Myasoedova, A. P. Athreya, C. S. Crowson, J. M. Davis III, K. J. Warrington, R. C. Walchak, E. Carlson, K. R. Kalari, T. Bongartz, P. P. Tak, et al. Toward individualized prediction of response to methotrexate in early rheumatoid arthritis: A pharmacogenomics-driven machine learning approach. *Arthritis Care & Research*, 74(6):879–888, 2022.

[12] S. Narkhede. Understanding confusion matrix, May 2019.

[13] M. C. Reid, C. Eccleston, and K. Pillemer. Management of chronic pain in older adults. *Bmj*, 350, 2015.

[14] R. Sangani. A comprehensive guide on model calibration: What, when, and how, Sep 2022.

[15] A. J. Steele, S. C. Denaxas, A. D. Shah, H. Hemingway, and N. M. Luscombe. Machine learning models in electronic health records can outperform conventional survival models

for predicting patient mortality in coronary artery disease. *PloS one*, 13(8):e0202344, 2018.

[16] H. Westerlind, M. Maciejewski, T. Frisell, S. A. Jelinsky, D. Ziemek, and J. Askling. What is the persistence to methotrexate in rheumatoid arthritis, and does machine learning outperform hypothesis-based approaches to its prediction? *ACR Open Rheumatology*, 3(7):457–463, 2021.