Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–17
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076241309505
journals.sagepub.com/home/dhj
Sage

Preliminary study: Data analytics for predicting medication adherence in Malaysian arthritis patients

Firdaus Aziz[1,*], Shubathira Sooriamoorthy[2,3,†], Bryan Liew[4,†], Sharifah M. Syed Ahmad[5], Wei Wen Chong[2], Sorayya Malek[4,*] and Adliah Mhd Ali[2,*]

Abstract

Objective: In multi-ethnic Malaysian populations, understanding and improving medication adherence in arthritis patients is crucial for enhancing treatment outcomes. Non-adherence, whether intentional or due to complex factors, can lead to severe long-term consequences such as increased disability and disease progression. This study analysed and predicted Malaysian arthritis medication adherence using 13 machine learning models.

Methods: A majority of 151 responders (82.1%) were female and 58.3% had comorbid illnesses. Notably, 90.07% of respondents were non-adherence to their prescription, with significant differences by occupation and aids in medication. This study's machine learning models perform better with recursive feature elimination for feature selection. Key variables included occupation, presence of other diseases, religion, income, medication aid, marital status, and number of medications taken per day. These variables were used to build predictive models for medication adherence.

Results: Results from machine learning algorithms showed varied performance. Support vector machine, gradient boosting, and random forest models performed best with AUC values of 0.907, 0.775, and 0.632 utilizing all variables. When using selected variables, random forest (AUC = 0.883), gradient boosting (AUC = 0.872), and Bagging (AUC = 0.860) performed best. Model interpretation using SHapley Additive exPlanations analysis identified occupation as the most important variable affecting medication adherence. The study also found that unemployment, concomitant disease, income, medication aid type, marital status, and daily medication count are connected with non-adherence.

Conclusion: The findings underscore the multifaceted nature of medication adherence in arthritis, highlighting the need for personalized approaches to improve adherence rates.

Keywords

Medication adherence, arthritis, machine learning, data analytics, Malaysia

Submission date: 3 September 2024; Acceptance date: 9 December 2024

[1] Pusat Pengajian Citra Universiti (School of Liberal Studies), Universiti Kebangsaan Malaysia, Bangi, Malaysia
[2] Center for Quality Management of Medicines (QMM), Faculty of Pharmacy, Universiti Kebangsaan Malaysia, Kuala Lumpur, Malaysia
[3] Pharmacy Department, Tengku Ampuan Rahimah Hospital, Klang, Malaysia
[4] Bioinformatics Science Programme, Institute of Biological Sciences, Universiti Malaya, Kuala Lumpur, Malaysia
[5] Faculty of Engineering, Universiti Putra Malaysia, Serdang, Malaysia
[*] These authors contributed equally to this work.
[†] These authors also contributed equally to this work.

Corresponding authors:
Sorayya Malek, Bioinformatics Science Programme, Institute of Biological Sciences, Faculty of Science, Universiti Malaya, Lembah Pantai, Wilayah Persekutuan, 50603 Kuala Lumpur, Malaysia.
Email: sorayya@um.edu.my

Adliah Mhd Ali, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Wilayah Persekutuan Kuala Lumpur, Malaysia.
Email: adliah@ukm.edu.my

Introduction

Arthritis is a chronic disease that causes one's joints to inflame and swell, and more than 350 million people were affected by arthritis globally.[1] Osteoarthritis was found to be the most prevalent type compared to the others.[2] Most types of arthritis do not directly cause mortality immediately; however, the condition may deteriorate and worsen patients' conditions and putting them at risk. Arthritis can be treated with medications such as anti-inflammatories or steroids, treatment, or surgery, depending on the type and severity.[3]

To effectively treat arthritis, patients must adhere to their medication regimens, also known as medication adherence. Failure to adhere to medication or non-adherence may worsen the condition resulting in severe long-term consequences, such as renal failure in lupus patients or flares and increased disability in rheumatoid arthritis (RA) patients.[4] Medication adherence, as defined by the World Health Organization, refers to the extent to which an individual's actions align with the prescribed guidelines provided by a healthcare professional.[5] Adherence refers to the collaborative effort between the patient and physician to enhance the patient's health by incorporating the physician's medical advice with the patient's lifestyle, values, and care preferences.[6] Non-adherence to medication occurs when a patient does not start, starts late, does not follow their dose schedule, or stops too soon.[7,8] Non-adherence can also come from patient rejection or unintended factors including socioeconomic concerns, medical issues, or treatment complexity.[9]

In Malaysia, data has shown that non-adherence among chronic disease patients was around 50% with hypertension and hyperlipidaemia as significant co-morbidities in almost 70% of the arthritis patients.[10] The treatment goal and approach should focus on providing symptom relief to effectively manage the patient's uncomfortable or distressing symptoms. The pharmacological approach to treating different types of arthritis varies based on the condition.[11] For osteoarthritis, first-line treatments include NSAIDs and acetaminophen for pain relief, with more severe cases potentially requiring corticosteroid injections or hyaluronic acid. In RA, disease-modifying anti-rheumatic drugs (DMARDs) such as methotrexate are the standard treatment, with biologic agents like TNF inhibitors added for patients with more severe or unresponsive symptoms. Similarly, in psoriatic arthritis, DMARDs and biologics are the primary treatments, with NSAIDs used for milder symptoms. Biologic therapies that target specific inflammatory pathways are often prioritized in more advanced cases of RA and psoriatic arthritis. Managing arthritis requires a long-term commitment as patients negotiate the difficulties of symptomatic treatment. That is why it is important to investigate the complex perspectives and beliefs of arthritis patients facing the continuous demands of arthritis management.

The cause of non-adherence in these patients was also due to complexity in taking multiple medications and in certain cases due to drug-related problems.[12] Understanding the cause of non-adherence is crucial. Sociodemographic information, patient factors such as drug beliefs and attitudes, diseases, therapies, the doctor–patient relationship, and social healthcare context interaction were found to be among the determining factors.[13]

Patients' perceptions about medicines were found to be one of the most well-studied factors that lead to non-adherence to medications.[14] Patients' understanding of their health condition and therapy may affect their belief about disease and treatment.[12] The Beliefs about Medicines Questionnaire (BMQ) was found to influence non-adherence.[15] The BMQ reflects chronic disease patients' medication perception.[5] The questionnaire explores general view of medications and beliefs and divides medicine beliefs into general (Overuse and Harm) and specific (Necessity and Concerns) categories.

Conventional statistical analysis is commonly used to determine the adherence of arthritis patients.[16–18] Recently, machine learning has been utilized in the pharmaceutical field to extract information and uncover correlations in data, one of which is determining patient adherence to drugs and treatments for various disorders. Desai et al.[19] employed Logistic Regression to predict medication adherence in fibromyalgia patients, with an area under the ROC curve (AUC) of 0.62. Recurrent Neural Network also demonstrated high accuracy in predicting drug adherence in individuals with RA.[20] However, due to the black box nature of machine learning methods, integrating machine learning models in pharmaceutical datasets is difficult. To circumvent the black box characteristics of machine learning algorithms, explainable machine learning, the SHapley additive exPlanations (SHAP), a model-independent local explain ability, can be utilized.[21]

Yet, to the best of our knowledge, there are not many studies discussing the application of machine learning approaches to discover factors associated with medication adherence in arthritis patients in general. Adherence failure leads to poor clinical outcomes; therefore, identifying individuals who have a proclivity for poor adherence behaviour might lower the likelihood of hospitalization and severe health outcomes. It also has societal implications because incorrect medicine use reduces the quality of data available for drug development and enhancement efforts. As a result, the purpose of this study is to determine the factors affecting arthritis patients' medication adherence using feature selection, which will subsequently be used to construct a prediction model using machine learning techniques. SHAP will then be utilized to evaluate the prediction models to determine the variables that contributed to arthritis patients' medication adherence.

Method

Data collection and preparation

Patients aged 18 years and above and diagnosed with arthritis were invited to participate in this study. An explanatory statement was provided to the patient. Should the patient consent to participate, they will be required to sign the informed consent document. Patients who required approval from guardians or parents were unable to understand English or native language were excluded from this survey. This study was conducted in a tertiary hospital outpatient clinic in Kuala Lumpur, Malaysia, from October 2014 to October 2015. This research received ethical approval in accordance with the UKM Research Ethics Committee, under the approval code UKM 1.5.3.5/244/SPP/NF-025-2011. The data collected included demographic data, duration of medication taking for arthritis, other comorbid disease, total number of medications taken per day, aids in medication taking for arthritis, and counselling received for arthritis medications.

This study used validated questionnaires entitled Beliefs about Medicines Questionnaire (BMQ) by Horne et al.[22] and Malaysian Medication Adherence Scale (MALMAS) by Chua et al.[23] The questionnaire consisted of three parts. The first section contains information about the patients' demographics, including their age, gender, ethnicity, degree of education, occupation, monthly income, and marital status. The second part of the questionnaire consisted validated BMQ to evaluate patients' beliefs on medication.[22] The BMQ questionnaire is divided into two sections: the BMQ-Specific, which evaluates views about medications used to treat a specific disease, and the BMQ-General, which evaluates beliefs about medications generally. These items reflect the individual views of the patients regarding the medications that are being administered to them. The 18 BMQ items further divided beliefs about medication use into Specific Necessity, Specific Concerns, General Harm, and General Overuse beliefs. General Overuse refers to the idea of over-prescription by physicians who place an excessive amount of confidence in medication, and General Harm evaluates ideas about how damaging medications are (harm scale). Each scale has four elements, and the possible total scores range from 4 to 20. Higher scores revealed more generalized negative opinions about drugs. The Specific Concerns scale is used to determine the possibility of negative consequences from taking the prescribed medication. The Specific Necessity scale investigates the patient's perception of their unique needs for sticking to their medication regimen. There are five items on each scale. Total scores ranged from 5 to 25 based on the sum of the individual item scores.[22] Higher scores in the General Harm and General Overuse categories indicate a negative perception of the medication. Similarly, higher scores obtained in the Specific Concerns category signify that adverse reactions are believed to be possibly harmful with regular intake of medication. Higher scores in the Specific Necessity category indicate the patient's need to adhere to medication to preserve good health.[22] The third section of the questionnaire consisted of patients' self-reported medication adherence as measured by the eight-item in MALMAS. The responses in the MALMAS had a total score which ranged from 0 to 8. Patients were considered adherent if they have a total score of 6 and above (≥6) and non-adherent if they have a total score of below 6 (<6).[23] We followed the adherence scoring and cut-off defined by the MALMAS study, a standard adherence framework for the Malaysian population, to maintain consistency and comparability with prior research.

Convenience sampling method was used in this study. Patients with arthritis disease were invited to participate in the study while waiting for their clinic appointment or medicines at the pharmacy counter. Patients who agreed to participate will answer the survey form. Patient who has been diagnosed with arthritis for at least 1 year with the minimum of one arthritis medication, adult patients, and those willing to cooperate were included in this study. Patients with terminal stage of disease, pregnancy with arthritis, immune-comprised patients, and having difficulty in communication were excluded from this study.

There were 16 variables used in this study, and the output parameter is adherence level. The data are collected via questionnaires and are used to calculate the adherence level among the arthritis patients. Patient's belief about medications may play an important role in determining patient's adherence to their treatment among arthritis patients. Belief in medications may vary depending on patients thought. Some patients may have different belief on medications due to the advantages and the suitability of medications, experience-related side effects, the effect of the medications, medical costs, type of diseases, psychological factors, demographic factors, and others.[24]

This study implemented categorical and continuous variables, which are summarized in Table 1. The questionnaire yielded three categories of measured variables: demographic characteristics of the patient, medication and disease history, and beliefs regarding medications.

Sample pre-processing

We developed machine learning models with a complete dataset and removed any missing values to ensure the validity of the results. A total of 151 respondents were obtained from the validated questionnaire and were identified as complete cases (with no missing values on predictors and outcome variables). Each variable was then subjected to a near-zero-variance or zero-variance test to eliminate those not exhibiting any variation to avoid sampling errors and unexpected outcomes. This rendered a full predictor set of

Table 1. Statistical analysis of variables used in the study.

Variables	Attributes	Total	Adherent	Non-adherent	*p*-value
N		151	15 (9.93)	136 (90.07)	
Age		57.8 ± 12.9	52.4 ± 14.5	58.43 ± 12.6	0.085[a]
Gender	Male	27 (17.9)	1 (6.7)	26 (19.1)	0.232[b]
	Female	124 (82.1)	14 (93.3)	110 (80.9)	
Religion	Islam	77 (51.0)	9 (60.0)	68 (50.0)	0.266[b]
	Buddhist	48 (31.8)	4 (26.7)	44 (32.4)	
	Hindu	7 (4.6)	2 (13.3)	5(3.7)	
	Christian	17 (11.3)	0 (0.00)	17 (12.5)	
	Others	2 (1.3)	0 (0.00)	2 (1.5)	
Occupation	Agricultural	1 (0.7)	0 (0)	1 (0.7)	0.028[b]
	Business	4 (2.6)	1 (6.7)	3 (2.2)	
	Education	14 (9.3)	5 (33.3)	9 (6.6)	
	Health	3 (2.0)	0 (0)	3 (2.2)	
	Housework	37 (24.5)	5 (33.3)	32 (23.5)	
	Engineering	4 (2.6)	0 (0)	4 (2.9)	
	Unemployed	59 (39.1)	3 (20.0)	56 (41.2)	
	Retiree	29 (19.2)	1 (6.7)	28 (20.6)	
Ethnicity	Malay	73 (48.3)	9 (60.0)	64 (47.1)	0.711[b]
	Chinese	66 (43.7)	5 (33.3)	61 (44.9)	
	Indian	8 (5.3)	1 (6.7)	7 (5.1)	
	Others	4 (2.6)	0 (0)	4 (2.9)	
Highest educational level	Primary	43 (28.5)	4 (26.7)	39 (28.7)	0.096[b]
	Secondary	63 (41.7)	4 (26.7)	59 (43.4)	
	Tertiary	45 (29.8)	7 (46.7)	38 (27.9)	
Monthly income	No income	5 (3.3)	0 (0)	5 (3.7)	0.828[b]
	<MYR1000	82 (54.3)	7 (46.7)	75 (55.1)	
	MYR1000–1999	12 (7.9)	2 (13.3)	10 (7.4)	
	MYR2000–2999	25 (16.6)	3 (20.0)	22 (16.2)	

(continued)

Table 1. Continued.

Variables	Attributes	Total	Adherent	Non-adherent	*p*-value
	MYR3000–3999	0 (0)	0 (0)	0 (0)	
	MYR4000–4999	23 (15.2)	3 (20.0)	20 (14.7)	
	>MYR5000	4 (2.6)	0 (0)	4 (2.9)	
Marital status	Married	113 (74.8)	11 (73.3)	102 (75.0)	
	Single/widow/widower	38 (25.2)	4 (26.7)	34 (25.0)	
Duration of medication intake	<1 year	22 (14.6)	4 (26.7)	18 (13.2)	0.265[b]
	1–4 years	58 (38.4)	7 (46.7)	51 (37.5)	
	5–10 years	37 (24.5)	3 (20.0)	34 (25.0)	
	>10 years	34 (22.5)	1 (6.7)	33 (24.3)	
Presence of other concomitant disease	Yes	88 (58.3)	12 (80.0)	76 (55.9)	0.072[b]
	No	63 (41.7)	3 (20.0)	60 (44.1)	
Total medicine taken per day	1	46 (30.5)	6 (40.0)	40 (29.4)	0.374[b]
	2	101 (66.9)	8 (53.3)	93 (68.4)	
	3	4 (2.6)	1 (6.7)	3 (2.2)	
Aid in medication	Yes	107 (70.9)	7 (46.7)	100 (73.5)	0.030[b]
	No	44 (29.1)	8 (53.3)	36 (26.5)	
General harm		10.2 ± 2.4	10.6 ± 2.0	10.13 ± 2.5	0.480[a]
General overuse		13.1 ± 2.2	13.3 ± 2.8	13.0 ± 2.2	0.688[a]
Specific necessity		17.4 ± 3.4	16.3 ± 2.9	17.5 ± 3.4	0.172[a]
Specific concern		16.6 ± 3.2	16.6 ± 3.8	16.6 ± 3.2	0.997[a]

Note. [a]Independent *t*-test. [b]Chi-squared test. *p*-value is statistically significant as $p < 0.05$. Data are expressed as count (percentage), or mean ± standard deviation, as appropriate.

16 variables (5 continuous, 11 categorical) for the study as shown in Table 1.

To prevent data leakage, all pre-processing steps were applied separately to the training and validation datasets to maintain the integrity of model evaluation.[25] Following the methodology outlined by Kuhn and Johnson,[26] we used stratified random sampling based on the outcome variable to ensure proportional representation of both adherent and non-adherent groups in each subset. Given that non-adherence is rare (only 15 individuals), this approach ensured that both the training and validation datasets had a balanced distribution of outcomes. The entire dataset of 151 respondents was split into 70% for model development (training dataset) and 30% for performance evaluation (validation dataset). Importantly, the validation set remained untouched during the training process and was used exclusively to assess model performance across time frames, ensuring unbiased evaluation.

For continuous variables in this study, such as age, specific concern, specific necessity, general overuse, and general harm, we used standardization, also known as *z*-score normalization. Data normalization is the process of converting continuous variable values in a dataset to a similar scale while keeping differences in value ranges.

Since our study dataset can be categorized as small, it is crucial to adopt robust methodologies to ensure the validity and reliability of our findings. A study by Nti et al.[27] suggests using a resampling method, k-fold cross-validation, which involves dividing the dataset into k subsets and training the model k times, each time using a different fold for validation. This method maximizes data utilization, reduces overfitting risk, and provides a comprehensive performance evaluation. In this study, 5-fold cross-validation was applied. To further ensure validity, we proposed using a separate validation set comprising 30% of the dataset. This independent validation set offers an unbiased performance assessment, aids in model selection and hyperparameter tuning, and estimates real-world performance. Combining k-fold cross-validation with a separate validation set ensures our study's findings are both reliable and generalizable, addressing the challenges associated with small datasets.

Machine learning algorithms such as Support Vector Machine,[28] Gaussian Naïve Bayes,[29] Logistic Regression,[30] Ridge Classifier.[31] Random Forest,[32] Bagging,[33] Gradient Boosting,[34] Linear Support Vector Classifier,[35] AdaBoost,[36] k-Nearest Neighbour,[37] Decision Tree,[38] Bernoulli Naïve Bayes,[39] and Gaussian Process Regression[40] were used to develop the prediction model for the medication adherence in this study in Python.

In this study, both a linear and radial basis function kernel were employed in conjunction with the robust learning methods, Support Vector Machine. They are statistical techniques for categorizing data that is difficult to separate linearly by mapping it to a large feature space. Simple classification methods like Naïve Bayes enable the quick creation of classification models. A type of simple 'probabilistic classifier' in machine learning, Naïve Bayes classifiers, makes the strong (naive) assumption that the features are conditionally independent. While Bernoulli Naïve Bayes assumes that the features are conditionally independent given the class label and follow a Bernoulli distribution, Gaussian Naïve Bayes assumes that the probability distribution of each feature given the class label is Gaussian. The k-Nearest Neighbour method uses 'feature similarity' or 'nearest neighbours' to determine which cluster a new data point belongs to. The number of neighbours it finds that are closest to a particular one gives the k-Nearest Neighbour its name.

Decision Tree is a non-parametric supervised learning method that is applied to regression and classification. Decision Tree serves as the main classifier while Random Forest uses bagging to produce numerous small decision trees. The class that Random Forest trees predicted to receive the most votes is used in the models. Gradient Boosting and AdaBoost are machine learning techniques that combine several weak learners (decision trees) to create a strong predictive model by successively training a model on a random sample of data. The training of weak learners, the loss functions they optimize, and the weight updates are where Gradient Boosting and AdaBoost diverge most. AdaBoost gives misclassified cases more weight, while Gradient Boosting focuses on reducing gradients in the loss function. Bagging, also known as Bootstrap Aggregating, is a machine learning technique that combines the predictions of various separately trained models to increase the stability and accuracy of models. The first step in bagging is to divide up the training data into various subgroups using replacement sampling. Each subset is the same size as the first training set, with the possibility of some samples being repeated and others being left out. The method is referred to as bootstrap sampling.

The Logistic Regression machine learning model is a member of the family of supervised machine learning models. For statistical analysis, the Logistic Regression technique is employed when the dependent variable is dichotomous. The link between one dependent binary variable and one or more independent nominal, ordinal, or ratio-level variables is defined and explained by the Logistic Regression approach. Ridge Classifier is a machine learning algorithm that is based on Ridge Regression, a regularization method used for regression issues. Ridge Regression is also known as Ridge Regression for classification. The primary principle of the Ridge Classifier is to reduce the influence of irrelevant features and avoid overfitting by using a linear model (usually logistic regression) with L2 regularization (Ridge regularization). It does this by including a penalty term in the loss function that pushes the model to strike a compromise between fitting the training set of data and minimizing the size of the model coefficients. Gaussian regression, commonly referred to as Gaussian Process Regression, is a probabilistic regression technique used in machine learning that uses Gaussian processes to explain the connection between input features and output values. It makes it possible to model complex functions in a flexible and non-parametric way. Random Forest, Support Vector Machine, Decision Tree, Naive Bayes, and XGBoost are all examples of machine learning algorithms that have been used to classify or predict adherence to medications.[41–44] In the study by Delpino et al.,[45] it was discovered that k-Nearest Neighbour, Naive Bayes, and Random Forest were some of the most often used machine learning algorithms for getting the best model performance in evaluating medication adherence.

Feature selection and model interpretation

After all the machine learning models had been developed using all the variables, the feature selection approach was used to identify the significant variables that influence the medication adherence among patients with arthritis. This study adopted recursive feature elimination as the feature selection method.[46] A greedy optimization technique to find the feature subset with the best performance is how recursive feature elimination works. It frequently produces models and stores the highest or least important characteristics for each iteration. Using the remaining features while they are still available, it constructs the subsequent model. After that, the features are sorted in reverse order of elimination. The features with the highest ratings are kept, while those with the lowest ratings are dropped. Recursive feature elimination is a type of feature selection that starts with every feature and continues in this manner without going backwards. This process is repeated numerous times until the necessary number of characteristics is either obtained or it is determined that the performance cannot be further enhanced. All the models were afterwards retrained using just the selected variables from the recursive feature elimination process. To prevent overfitting for model building on the training set, 5-fold cross-validation was performed. Then, the model with the best calibration and discrimination values was chosen as the best-performing model. Feature selection in machine learning speeds up model building and improves algorithm computational efficiency and saving time.

However, it is challenging to use machine learning models in pharmaceutical contexts because of their 'black box' character. Machine learning models are agnostic; thus, changing the input and viewing the results can show how the underlying model behaves.

The input can be interpreted by altering components that are comprehensible to humans. As a result, this study used SHAP to analyze the best machine learning model in this research. The predictors are ranked by significance using the SHAP technique, with the first predictor being the most significant and the last being the least significant. By comparing the relative magnitude of the mean absolute SHAP values for the various features, the SHAP algorithm may be used to assess the significance of each feature. The SHAP beeswarm plot shows an overview of how the dataset's chosen factors affect the model's output. The amount and direction of each feature that affects anticipated adherence to medications are quantified by the SHAP values. On each feature row, a single dot stands in for the explanation in each case. Thus, each little dot on the diagram corresponds to a single observation or piece of data. The original value of a feature is shown in colour in the SHAP plots. Red data points show high feature values, whereas blue data points represent low feature values.[21]

Model evaluation, validation, and performance measures

The calibration of the model was assessed using standardized metrics. As a class-insensitive predictive performance indicator, the AUC was employed. Accuracy, recall, and precision score were additional performance indicators for model calibration. A paired resampled *t*-test was employed to examine the prediction capabilities of the different machine learning models.

Additional statistics

In this study, categorical variable frequencies and continuous variable means and standard deviations are reported. Variable relationships were identified through correlation analysis. A chi-square test was employed in univariate analysis to identify significant variables, and a two-sided independent Student's *t*-test ($p < 0.05$) was utilized to compare them. To assess the effectiveness of machine learning, pairwise corrected resampled *t*-tests were utilized. Figure 1 illustrates the model development process in this study.

Results

Respondents' characteristics

A total of 151 respondents participated in this study. Table 1 shows the patients' characteristic used in this study. The mean age was 57.8 (SD = 12.0) and majority (with 82.1%) of the respondents were female; 58.3% have other comorbid disease. Of the total respondents, 90.07% did not adhere to their medication. Significant differences were noted between respondents who adhere and not adhere to their medications in terms of occupation and aids in medication ($p < 0.05$).

Feature selection

This study adopted recursive feature elimination as a feature selection method in reducing the number of variables to increase the performance of the machine learning models. The variables selected through the feature selection method are occupation, presence of other concomitant diseases, religion, income, aid in medication, marital status, and number of medications taken per day. The selected variables were then implemented into each of the algorithm. Hence, there are two sets of data being used for each of the algorithm to build the machine learning model in predicting the medication adherence in patients with arthritis. Table 2 shows the performance of the machine learning models using all set of variables (16 variables) while Table 3 shows the performance of the machine learning models using the selected variables (7 variables).

Model performance

Based on Table 2, Support Vector Machine, Gradient Boosting, and Random Forest are the top performing machine learning models with all variables using the 30% untouched validation dataset. The Support Vector Machine algorithm outperforms Gradient Boosting; however, there is no statistically significant difference in their performance ($p = 0.317$). However, there is a substantial difference in performance when compared to Random Forest ($p = 0.018$). The performance of Gradient Boosting

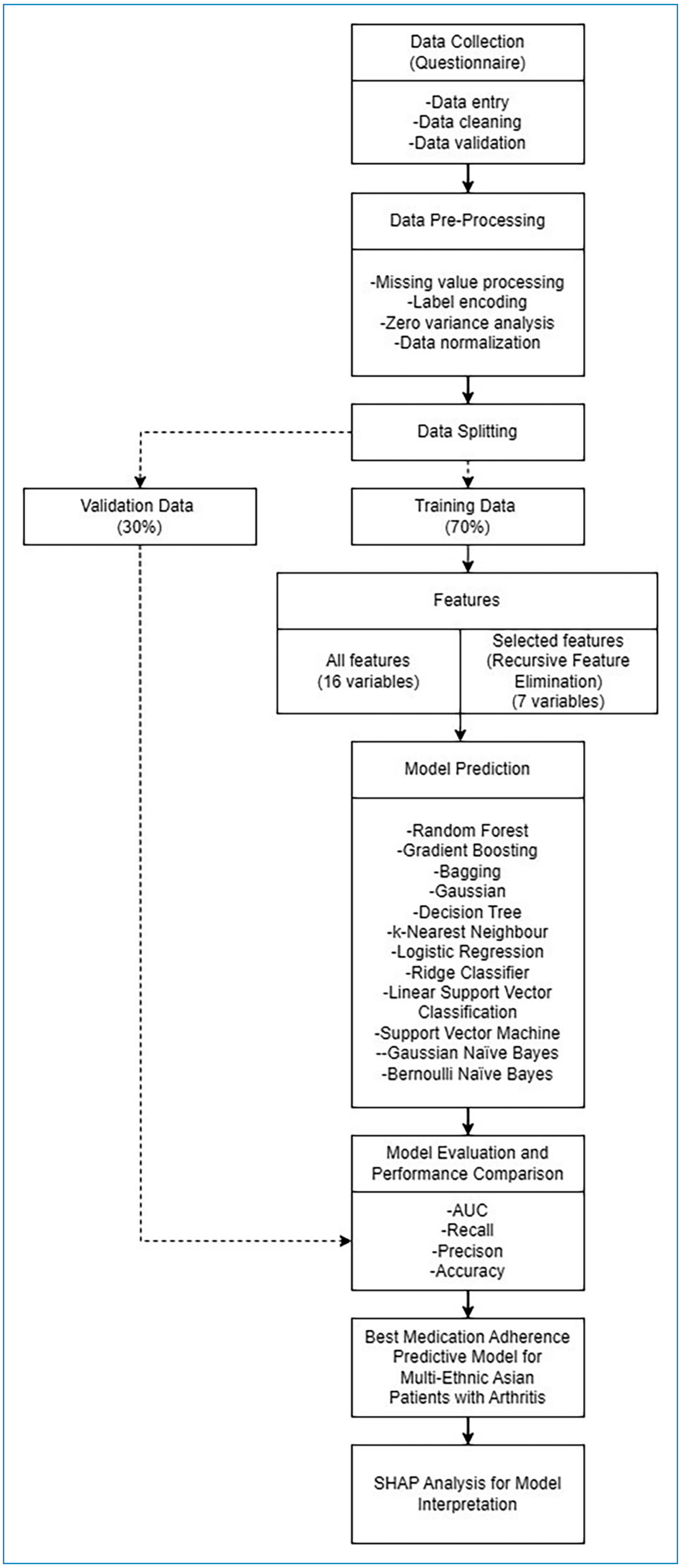

Figure 1. Flowchart of the predictive model development.

Table 2. The performance of the machine learning algorithms using all variables based on the 30% untouched validation dataset.

Machine learning algorithm	AUC (95% CI)	Recall	Precision	Accuracy
Random Forest	0.632 (0.801, 0.981)	0.333	0.250	0.891
Gradient Boosting	0.775 (0.772, 0.967)	0.667	0.286	0.870
Bagging	0.609 (0.744, 0.951)	0.333	0.167	0.848
Gaussian	0.430 (0.690, 0.920)	0.001	0.001	0.804
Decision Tree	0.573 (0.663, 0.901)	0.333	0.111	0.783
k-Nearest Neighbour	0.504 (0.515, 0.790)	0.333	0.667	0.652
Logistic Regression	0.612 (0.422, 0.708)	0.667	0.095	0.565
Ridge Classifier	0.612 (0.422, 0.708)	0.667	0.095	0.565
Linear Support Vector Classification	0.554 (0.313, 0.600)	0.667	0.077	0.457
Support Vector Machine	0.907 (0.716, 0.935)	1.000	0.272	0.826
AdaBoost	0.694 (0.587, 0.848)	0.667	0.143	0.717
Gaussian Naïve Bayes	0.659 (0.515, 0.790)	0.667	0.118	0.652
Bernoulli Naïve Bayes	0.601 (0.400, 0.687)	0.667	0.091	0.543

Table 3. The performance of the machine learning algorithms using selected variables based on the 30% untouched validation dataset.

Machine learning algorithm	AUC (95% CI)	Recall	Precision	Accuracy
Random Forest	0.883 (0.645, 0.920)	1.000	0.231	0.783
Gradient Boosting	0.872 (0.623, 0.899)	1.000	0.214	0.761
Bagging	0.860 (0.601, 0.877)	1.000	0.200	0.739
Gaussian	0.849 (0.580, 0.855)	1.000	0.188	0.717
Decision Tree	0.849 (0.580, 0.855)	1.000	0.188	0.717
k-Nearest Neighbour	0.837 (0.558, 0.833)	1.000	0.176	0.696
Logistic Regression	0.802 (0.493, 0.768)	1.000	0.150	0.630
Ridge Classifier	0.802 (0.493, 0.768)	1.000	0.150	0.630
Linear Support Vector Classification	0.802 (0.493, 0.768)	1.000	0.150	0.630
Support Vector Machine	0.767 (0.428, 0.703)	1.000	0.130	0.565
AdaBoost	0.671 (0.536, 0.811)	0.667	0.125	0.674
Gaussian Naïve Bayes	0.647 (0.493,0.768)	0.667	0.111	0.630
Bernoulli Naïve Bayes	0.531 (0.275,0.550)	0.667	0.071	0.413

AUC: area under the ROC curve.

and Random Forest is significantly different ($p = 0.044$). In Table 3, the machine learning models that performed the best with selected variables (using feature selection) and the 30% untouched validation data are Random Forest, Gradient Boosting, and Bagging. There is no significant difference in performance between Random Forest and Gradient Boosting ($p = 0.183$), and Bagging ($p = 1.000$). It also shows that Gradient Boosting is performing significantly better when compared with Bagging ($p = 0.261$).

Feature selection is a strategic process in machine learning that involves carefully choosing the most important variables from a pool of potential factors, not only reducing the number of variables to enhance computational efficiency but also diminishing the risk of overfitting, ultimately making the model more cost-effective and boosting overall performance. Therefore, when the seven selected variables (selected using recursive feature elimination) are incorporated in the machine learning models, the performance of certain models improves while others deteriorate relative to the models' performances utilizing all variables, as evidenced by the AUC values mentioned above. The Support Vector Machine, AdaBoost, and Gaussian Naïve Bayes machine learning models demonstrate a decline in performance when only selected variables are used in the model building. Conversely, the remaining models suggest an improvement in performance. The Random Forest model, when utilizing selected variables, achieves the better AUC value as compared when using all variables. The difference in performance is statistically significant with $p = 0.018$. Support Vector Machine model with all variables has the highest predictive AUC value compared to Random Forest model with selected variables; however, the difference is not significant ($p = 1.000$). Thus, Random Forest model with selected variables is chosen as the best model since it has the least number of variables and the best predictive performance.

SHAP analysis

In order to properly understand the predictions made by the machine learning model, SHAP analysis was employed. This study utilized the knowledge gained from the best-performing model (Random Forest with selected variables), which was carefully selected and optimized. The diagram of the SHAP analysis is shown in Figure 2. The *y*-axis represents the variable name arranged in a descending order of significance, with occupation being the most significant, followed by the presence of concomitant disease, religion, income, aid in medication, marital status, and finally the total number of medications used each day.

The SHAP value is represented on the *x*-axis, while the gradient colour reflects the initial value of the variable. The orientation of the bar (left or right) for each characteristic indicates whether the effect is favourable or unfavourable. According to the SHAP analysis provided, the variable occupation exhibits the largest spread of SHAP values, indicating that it has a substantial influence on medication adherence. Higher values, represented by the colour red, appear to consistently correlate with a negative impact on adherence. This implies that individuals who are not unemployed or retired (based on Table 1) tend to exhibit low levels of adherence to their prescription. The presence of concomitant disease is associated with a distinct cluster of red dots on the positive side of the SHAP value axis. This indicates that having concomitant diseases may have a beneficial impact on adherence, presumably because individuals with these medical conditions are more mindful of their health. The relationship between income and

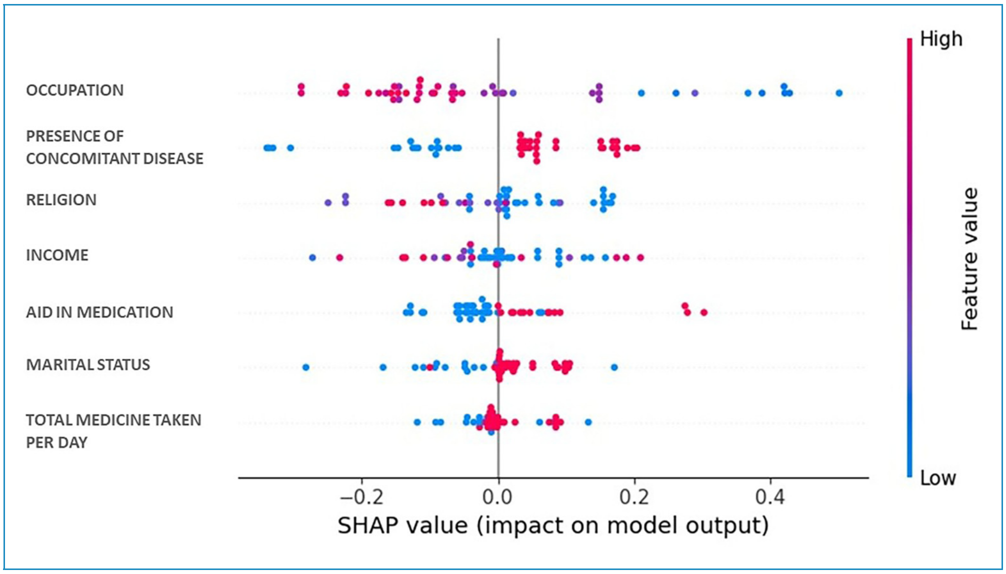

Figure 2. SHAP value for Random Forest model with selected variables.

adherence is inconclusive, as indicated by the presence of both red and blue dots dispersed around the zero line. This suggests that while higher income is associated with improved adherence for certain individuals, it does not exert a significant influence for others. There are many red dots to the right of the zero line, suggesting that having aid in medication using the traditional and modern way of using pillbox, timetable, and applications is generally associated with better adherence. The distribution of marital status reveals that the red dots are positioned to the right of the zero line, indicating that individuals who are married (higher values) exhibit greater adherence to their medications. Lastly, there is a slight spread, but many dots are mostly centred around zero but incline more towards the left of the zero line, which might indicate that total medicine taken per day has a slight negative effect on medication adherence.

Discussion

Evaluating the association between self-reported adherence and medicine belief can yield valuable insights into individuals' choices to adhere with their treatment regimen. This study is the first to utilize 13 machine learning models to forecast and identify the elements that influence adherence levels among Malaysian arthritis patients. The algorithms for machine learning were created utilizing variables linked to adherence to arthritis medicine that were determined by recursive elimination methods. The present study's findings indicate that (i) the performance of machine learning models was enhanced by utilizing features selected through recursive feature elimination and k-fold validation approach; (ii) the three best-performing machine learning models, Random Forest (AUC = 0.883), Gradient Boosting (AUC = 0.872), and Bagging (AUC = 0.860), outperform all other models when constructed using selected features; (iii) SHAP enables the evaluation of features related to the medication adherence of patients with arthritis.

In examining medication adherence among a multi-ethnic population in Malaysia, it is essential to consider the unique cultural, social, and economic factors that characterize this region. Cultural beliefs can significantly shape patients' views on medication efficacy and safety; for instance, some patients perceive modern medications as potentially harmful, adopting a fatalistic attitude towards treatment outcomes.[47] Additionally, economic factors play a critical role, as cost-sharing practices and limited access to medications may disproportionately affect those from lower socioeconomic backgrounds, who may prioritize immediate financial constraints over long-term health needs.[48] This economic disadvantage often amplifies perceived medication risks, leading to decreased adherence rates among economically marginalized groups.[49] Given these contextual influences, studying the medication beliefs of multi-ethnic arthritis patients is crucial to understanding adherence behaviours and identifying targeted strategies that can improve health outcomes across diverse populations.

Both direct and indirect techniques of measurement can be used to categorize conventional methods of evaluating drug adherence. One of the most accurate yet invasive methods is the direct approach, which measures drug concentrations or its metabolite in blood or urine.[50] Patient questionnaires, pill counts, prescription fill rates, clinical outcome assessment, and electronic medication monitors are examples of indirect approaches.[51] Every technique has advantages and disadvantages, and one could serve as the benchmark for a different strategy. The Morisky Medication Adherence Scale (MMAS) is a patient questionnaire that is widely utilized for evaluating medication adherence. MALMAS, on the other hand, was created in Malaysia and slightly modified to fit the environment here. This approach relies on patient recollection and could be biased depending on the patient's reaction. Utilizing a validated and well-structured questionnaire, previously assessed for reliability and validity, ensures that the questions are meticulously crafted to reduce the propensity for socially desirable responses.[52]

The accuracy of a medication adherence screening tool was examined in a systematic review and meta-analysis of the MMAS-8 by Moon et al.[53] The results showed that almost half of the studies examining diagnostic accuracy used a reference standard that was either highly risky or had unclear descriptions in terms of both the assessment of risk of bias and applicability. Despite reports that Cronbach's alpha is inappropriate for assessing a test's internal consistency and reliability, all the research in this study employed it. However, Moon et al. pointed out that while MMAS-8 is being developed, individual research that calculated sensitivity and specificity used the cut-off value of 6, as recommended by Morisky. More details regarding the MMAS-8's diagnostic accuracy may have been given if they had addressed sensitivity and specificity using various cut-off values. Since Morisky proposed 6 as the cut-off value in MMAS-8 during its development, most research report criteria validity results with 6 as the cut-off value. Since the accuracy of the assessment method employed above has flaws, it is best to apply alternate dichotomous methods such as machine learning.

There have been no studies on the impact of machine learning algorithms on predicting arthritis patients' adherence to treatment, though. The research gap has been filled by this work. Regression-based model using Artificial Neural Network was utilized by Aziz et al.[54] to predict the medication adherence level of hypertension patients using machine learning with root mean square error of 1.42. With an AUC of 0.62, a 2019 study by Desai et al.[19] used logistic regression to predict medication adherence in fibromyalgia patients. Eleven machine learning classifiers using Support Vector Machine predicted medication adherence with an AUC of 0.82 for patients with heart failure, according to Karanasiou et al.[44] Ten-fold stratified cross-validation and the synthetic minority over-sampling technique were employed in the study to balance the data. The model trained using basic Classification and Regression Trees has the highest prediction accuracy of all the classifiers they used, and they also proposed that predictive models with adequate accuracy can be used to better manage patients with heart failure and increase their adherence.

Son et al.[55] employed a support vector machine algorithm to predict the level of medication adherence in individuals suffering from heart failure. Due to the limited dataset consisting of only 76 patients, the study employed leave-one-out cross-validation. Their predictive model achieved an accuracy of 77.63%, and their feature selection analysis indicated that a longer term of diagnosis may result in non-adherence to medication, while a greater degree of education can enhance medication adherence. While Wu et al.[42] utilized 14 machine learning algorithms to construct a predictive model for examining the medication adherence of individuals with type 2 diabetes. The ensemble model yielded the highest AUC of 0.82 when employed to construct the prediction model. The accuracy of the model was validated by two-way cross-validation and evaluated using AUC. Mirzadeh et al.[56] employed machine learning techniques to predict adherence to medications in persons at elevated risk of atherosclerotic cardiovascular disease. Their study employed Classification and Regression Trees, Support Vector Machine, AdaBoost Regression, and Gradient Tree Boosting methods to train the predictive model. The models were then validated using 4-fold cross-validation. Their investigation indicated that all three of the algorithms employed, particularly the Support Vector Machine, are appropriate for predicting medication adherence among persons with a high susceptibility to atherosclerotic cardiovascular disease. Additionally, they acknowledged that factors such as understanding of medication, involvement, social assistance, empowerment, and self-assurance amid emergencies are crucial elements that impact adherence to medication. Overall, this research indicated that machine learning algorithms employing k-fold validation are effective approaches for patient prediction as this approach optimizes the utilization of small dataset, mitigates the risk of overfitting, and offers a thorough performance evaluation. Nevertheless, the AUC and accuracy obtained in this study surpass those of the previously published study that employed similar performance measures. The cross-validated AUC values provided impartial estimations of the performance of the machine learning classifiers trained with small data. It is crucial to recognize that when the sample size is limited, complex machine learning classifiers typically demonstrate reduced performance. Nevertheless, our study revealed that the machine learning models exhibited strong performance, despite the relatively small sample size of 151 individuals. This was particularly evident when utilizing selected variables, as opposed to employing all variables. The enhancement was evident in the cross-validated AUC values, demonstrating that satisfactory performance may still be attained with limited data.

The findings of this study suggest that machine learning algorithms utilizing Recursive Feature Elimination had superior performance compared to all other machine learning models developed utilizing the whole set of features, except for Support Vector Machine, AdaBoost, Gaussian Naïve Bayes, and Bernoulli Naïve Bayes. Feature selection algorithms play a crucial role in enhancing the effectiveness of machine learning algorithms.[57] Feature selection algorithms enhance the performance of machine learning models by lowering the dimensionality of predictors, therefore allowing for a suitable number of predictors to be used.[58] The study saw an improvement in the model's performance as the number of predictors decreased. This is because all the models containing selected variables exhibited significant clinical factors that contribute to the prediction of adherence. However, different case can be seen in Support Vector Machine as the AUC value decreases when less variables are used (selected variables from the recursive feature elimination). This may be due to Support Vector Machines that are susceptible to overfitting, particularly when working with a high number of features. When irrelevant or noisy features are used, the Support Vector Machine may overfit the training data by collecting noise rather than the underlying patterns. This further strengthens the choice of Random Forest with selected variables as the best model. The selected features in this study, listed in descending order of relevance, are occupation type, existence of concomitant disease, religion, income, aid in medication, marital status, and total medicine consumed per day.

Machine learning models are commonly referred to as black box models. This study utilized SHAP analysis to demonstrate that the association between input variables and clinical outcomes in the data may be elucidated. The study utilized SHAP to gain a deeper comprehension of the association between specific variables linked to adherence levels in patients with arthritis. Medication adherence is a multifaceted phenomenon with numerous interconnected factors. This study elucidates various characteristics correlated with patients' compliance to arthritis medications. Comprehending the connections between adherence and demographic elements is crucial due to the wide range and intricacy of medicine consumption patterns. The results of this study demonstrate that non-adherence relates to several factors, including unemployment, having a concurrent illness, having a high or low income, not using a timetable or pillbox as a medication aid, being unmarried, and taking a high total number of medications per day.

Unemployed individuals and pensioners frequently face difficulties in complying with their medicine regimes. Unemployment may result in difficulties affording necessary medications and healthcare costs, which can lead to inconsistent adherence.[59] The lack of a well-defined daily regimen, commonly found in employment, might lead to lapses in memory or inconsistent drug adherence. However, elderly individuals who receive pensions may face challenges in managing complex prescription schedules as a result of health issues associated with aging, despite having a stable income.[60] The elderly frequently face challenges in adhering to medication, which can be attributed to forgetfulness, limited comprehension of medication schedules, physical impairments in administering the prescription, or financial limitations.[61] Their capacity to appropriately handle medications may be further complicated by cognitive decline, social isolation, and mobility concerns. The complex nature of these difficulties underscores the significance of addressing not just economic concerns but also social and healthcare-related components to improve medication adherence among vulnerable people.

On the other hand, having several health disorders may increase patients' awareness of the significance of following prescribed therapies, as they acknowledge the interconnectedness of different elements of their health.[62] Within the realm of arthritis and related problems like cardiovascular disorders or diabetes, individuals may acquire a holistic comprehension of the interdependence of their well-being. The comprehensive viewpoint can operate as a compelling incentive for maintaining regular compliance with medication regimens. Furthermore, persons who are dealing with the intricacies of many health conditions may develop more robust relationships with healthcare practitioners, resulting in enhanced communication and collaborative decision-making regarding treatment strategies.[63] The implementation of a synergistic strategy in the management of several health disorders can potentially enhance patients' sense of empowerment, leading to a greater dedication to following recommended treatments for both arthritis and any concurrent illnesses.

Low-income patients may take medications sporadically because they find it difficult to pay for prescription drugs in addition to other necessities.[64] Anxiety about running out of scarce resources could lead to sporadic or incomplete adherence. However, even when they have access to prescription drugs, high-income people may not take them as directed because they believe their health is unaffected, resulting in irregular drug schedules.[65] Higher earners may also exhibit irregular adherence due to busy schedules and a propensity for alternative health approaches. The optimization of health outcomes in both cases still depends on resolving financial concerns, improving health literacy, and emphasizing the significance of consistent drug use.

Other than timetables and pillboxes, modern pharmaceutical aids improve adherence. Technology in mobile apps and smart medicine packaging reminds patients to take their medications on time and track adherence.[66] Personalized voice response systems and text message reminders engage patients through accessible communication channels. Pharmacist-led treatments and community support platforms promote accountability through personalized guidance and peer support.[67] Customized medicine programmes for lifestyle and preferences improve adherence.

This study also demonstrates that married patients more adhere to medications than single or divorced people. Spouse support is vital for reminding people to take their medicines.[68] Joint health management encourages shared responsibility for each other's health in married couples. Divorced or single individual may forget to take their medication because of a lack in this support structure. Marriage's emotional and practical support appears to improve medication adherence, underscoring the social dimension's impact on health behaviours.

When prescribed a high number of pills each day for chronic illnesses such as arthritis, adherence can be difficult. The intricacy of several medications, each with its own dosage and schedule, can cause disorientation and forgetfulness.[69] Arthritis often involves long-term treatment, and the cumulative effect of numerous medications may contribute to physical discomfort or unwanted side effects, potentially discouraging individuals from adhering to the prescribed regimen. High pill burdens can lower quality of life, driving intentional non-adherence as a coping tactic.[70] To improve adherence in arthritis, patients simplifying prescription regimens, providing clear instructions, and addressing cost and side effect concerns are crucial.

Limitations of the study

The primary limitation of our study is the relatively small sample size, consisting of only 151 respondents, to predict medication adherence in arthritis patients using machine learning. The main disadvantage of a small dataset is the potential for overfitting, where the model learns the training data too well, capturing noise instead of underlying patterns, thus performing poorly on new, unseen data. Additionally, small datasets may not capture the full variability and diversity of the target population, limiting the generalizability of the findings. However, small datasets can still be reliable if handled with appropriate techniques. Methods such as k-fold cross-validation, where the data is divided into k subsets and the model is trained and validated k times using different folds, maximize data utilization and reduce overfitting risks. Additionally, using an untouched validation set helps assess the model's performance on entirely unseen data, providing an unbiased performance estimate. Moreover, employing feature selection enhances the performance of the model, particularly in cases where the dataset is small. In such scenarios, the process of feature selection becomes even more crucial because selecting high-quality features can compensate for the limited amount of data by extracting the most informative aspects of the dataset.[71] There are also other studies that used small datasets but still able to give a good prediction using machine learning.[72–74] Combining these methods ensures robust model evaluation and enhances the credibility of findings derived from small datasets; however, in the future, we plan to conduct a more extensive study to gather additional data, with the aim of better representing the broader population of arthritis patients in Malaysia and across Asia. This effort aligns with our commitment to advancing this ongoing research.

In our study, the Random Forest model demonstrated relatively high accuracy with selected variables; however, its precision was limited by the imbalanced dataset, which included only 15 adherent patients out of 136 non-adherent patients. This limited number of adherent cases represents a study constraint. Altering the classification threshold could improve precision, but doing so would reduce comparability of our findings with the MALMAS adherence framework, which is based on real-world adherence patterns in the Malaysian population. When dealing with imbalanced outcomes, it is important to consider the clinical implications of prioritizing either accuracy or precision. Although the model achieved high recall by identifying most adherent patients, the low precision highlights the risk of false positives, which could result in unnecessary interventions, increased costs, and a potential decline in patient trust. These factors indicate the need to address precision-related issues before pursuing further development or clinical registration of the model.

Additionally, the clinical implications of the model's low precision are critical to consider. Misclassifying non-adherent patients as adherent could lead to delays in essential interventions, resulting in poor health outcomes and increased healthcare costs.[75] It could also lead to an underestimation of non-adherence, affecting resource allocation for targeted interventions. To ensure meaningful clinical impact, future work should focus on improving precision through expanded datasets, adjustments to prediction thresholds, and integration of additional variables, such as demographics, behavioural factors, and objective adherence measures like electronic medication monitors. Although the AUC performance on the validation set shows potential for clinical applicability, further data collection and model refinement are essential to establish the model as a reliable tool for healthcare providers.

Lastly, one other limitation of this study is the reliance on indirect methods to measure patient adherence, particularly self-reported data. Self-reported adherence is susceptible to overestimation, as patients may provide socially desirable responses or may lack accurate self-awareness regarding their medication-taking behaviour. This potential inflation in reported adherence rates can impact the validity of the findings, leading to an overly optimistic view of treatment compliance even though we have utilized a validated questionnaire designed to minimize bias and reduce the likelihood of patients either over- or under-claiming adherence. Nonetheless, interpretations should still be approached with caution, and future research could benefit from integrating more objective adherence measures to further enhance accuracy.

Conclusion

In conclusion, the study's novelty lies in its innovative application of machine learning algorithms in the pharmaceutical field, specifically targeting medication adherence – a domain where such applications are relatively unexplored. The research employed recursive feature elimination for feature selection, enhancing the performance of the machine learning models. SHAP analysis further explains the influence of these selected variables.

The study underscores the multifaceted nature of medication adherence, impacted by various factors including unemployment, concurrent illness, income level, medication aids, marital status, and the number of daily medications. Machine learning models, especially those utilizing Recursive Feature Elimination, proved effective in predicting medication adherence, outperforming models developed with the complete set of features. This novel approach not only enhances the analysis but also offers valuable insights for better patient management, paving the way for more personalized and effective healthcare interventions. This pioneering application of machine learning techniques in the field promises significant advancements in understanding and improving medication adherence.

Acknowledgements: We gratefully acknowledge the Ministry of Higher Education Malaysia for funding this study under the Fundamental Research Grant Scheme (FRGS/1/2021/SKK0/UKM/02/5), making this research possible. Our sincere appreciation also goes to the dedicated team at the tertiary hospital outpatient clinic in Kuala Lumpur, Malaysia, whose invaluable assistance in data collection and unwavering commitment were crucial to the successful completion of this study.

Contributorship: FA: data curation, formal analysis, investigation, methodology, software, validation, visualization and writing the original draft, and editing the final version of the manuscript. SS: data curation, formal analysis, methodology, validation, and writing the original draft. BL: data curation, formal analysis, investigation, software, and writing of the original draft of the manuscript. SMSA: conceptualization, funding acquisition, and reviewing the final version of the manuscript. CWW: conceptualization, funding acquisition, and reviewing the final version of the manuscript. SM: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration,

resources, software, supervision, validation, visualization, writing the draft, and also reviewing and editing the final manuscript. AMA: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing the draft, and also reviewing and editing the final manuscript. All authors read and approved the final manuscript.

Data availability: The datasets utilized and analysed in this study are available from the corresponding author upon reasonable request.

Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical approval: This research received ethical approval in accordance with the UKM Research Ethics Committee, under the approval code UKM 1.5.3.5/244/SPP/NF-025-2011.

Funding: The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ministry of Higher Education Malaysia under the Fundamental Research Grant Scheme (FRGS/1/2021/SKK0/UKM/02/5). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Guarantor: AMA.

Informed consent: Patients aged 18 years and above and diagnosed with arthritis were invited to participate in this study. An explanatory statement was provided to the patient. Should the patient consent to participate, they will be required to sign the informed consent document. Patients who required approval from guardians or parents were unable to understand English or native language were excluded from this survey. This study was conducted in a tertiary hospital outpatient clinic in Kuala Lumpur, Malaysia, from October 2014 to October 2015. This research received ethical approval in accordance with the UKM Research Ethics Committee, under the approval code UKM 1.5.3.5/244/SPP/NF-025-2011.

ORCID iD: Sorayya Malek https://orcid.org/0000-0001-6450-6404

Supplemental material: Supplemental material for this article is available online.

References

1. Callahan LF, Cleveland RJ, Allen KD, et al. Racial/ethnic, socioeconomic, and geographic disparities in the epidemiology of knee and hip osteoarthritis. *Rheumatic Disease Clin* 2021; 47: 1–20.
2. Dahl-Popolizio S, Smith K, Day M, et al. (eds) *Primary care occupational therapy: a quick reference guide*. Gewerbestrasse, Switzerland: Springer, 2023.
3. Jain BV, Pawar SR, Sabe AA, et al. Impact of OTC purchase and utilization of pain killers in rheumatoid arthritis. *J Surv Fish Sci* 2023; 10: 1062–1069.
4. Gong G, Dong A, Zhang Z, et al. Medication adherence and predictive factors among patients with rheumatoid arthritis: a COM-B model guided structural equation modeling analysis. *Patient Educ Couns* 2024; 119: 108080.
5. Jimmy B and Jose J. Patient medication adherence: measures in daily practice. *Oman Med J* 2011; 26: 155–159.
6. Rathbone AP, Mansoor SM, Krass I, et al. Qualitative study to conceptualise a model of interprofessional collaboration between pharmacists and general practitioners to support patients' adherence to medication. *BMJ Open* 2016; 6: e010488.
7. Gast A and Mathes T. Medication adherence influencing factors – an (updated) overview of systematic reviews. *Syst Rev* 2019; 8: 1–7.
8. Hebing RC, Aksu I, Twisk JS, et al. Effectiveness of electronic drug monitoring feedback to increase adherence in patients with RA initiating a biological DMARD: a randomised clinical trial. *RMD Open* 2022; 8: e001712.
9. Hebing RC, Elhendy N, van Geel EH, et al. The correlation between 4 adherence measurements methods in patients with rheumatoid arthritis using methotrexate. *Br J Clin Pharmacol* 2024; 90: 882–889.
10. Mahmood S, Jalal Z, Hadi MA, et al. Prevalence of non-adherence to antihypertensive medication in Asia: a systematic review and meta-analysis. *Int J Clin Pharm* 2021; 43: 486–501.
11. CPG Arthritis (Academy of Medicine, Malaysia/Clinical Practice Guidelines/Rheumatology), https://www.acadmed.org.my/index.cfm?menui d=67#Rheumatology (accessed 24 October 2024).
12. Pharmaceutical Services Division Ministry of Health, Malaysia. *Protocol medication therapy adherence clinic: rheumatology*. 1st ed. Kuala Lumpur: Perpustakaan Negara Malaysia, 2017.
13. Balsa A, de Yébenes MJ and Carmona L. Multilevel factors predict medication adherence in rheumatoid arthritis: a 6-month cohort study. *Ann Rheum Dis* 2022; 81: 327–334.
14. Nie B, Chapman SC, Chen Z, et al. Utilization of the beliefs about medicine questionnaire and prediction of medication adherence in China: a systematic review and meta-analysis. *J Psychosom Res* 2019; 122: 54–68.
15. Hannech E, Boussaid S, Rekik S, et al. Belief and adherence in rheumatoid arthritis patients in biological drugs. *BMJ Annal Rheumatic Dis* 2022; 81: 1310.
16. Smolen JS, Gladman D, McNeil HP, et al. Predicting adherence to therapy in rheumatoid arthritis, psoriatic arthritis or ankylosing spondylitis: a large cross-sectional study. *RMD Open* 2019; 5: e000585.
17. Turcu-Stiolica A, Mihaela-Simona S, Paulina Lucia C, et al. The influence of socio-demographic factors, lifestyle and psychiatric indicators on adherence to treatment of patients with rheumatoid arthritis: a cross-sectional study. *Medicina (B Aires)* 2020; 56: 178.
18. Nestoriuc Y, Orav EJ, Liang MH, et al. Prediction of nonspecific side effects in rheumatoid arthritis patients by

beliefs about medicines. *Arthritis Care Res (Hoboken)* 2010; 62: 791–799.

19. Desai R, Jo A and Marlow NM. Risk for medication nonadherence among Medicaid enrollees with fibromyalgia: development of a validated risk prediction tool. *Pain Pract* 2019; 19: 295–302.
20. Margffoy-Tuay EA, García-Hernandez C and Solano-Beltrán DC. Medication adherence improvement on rheumatoid arthritis patients based on past medical records. In: 2018 IX International Seminar of Biomedical Engineering (SIB), Bogota, Colombia, 16–18 May 2018. IEEE, pp. 1–6.
21. Kanyongo W and Ezugwu AE. Feature selection and importance of predictors of non-communicable diseases medication adherence from machine learning research perspectives. *Inf Med Unlocked* 2023; 38: 101232.
22. Horne R, Weinman J and Hankins M. The beliefs about medicines questionnaire: the development and evaluation of a new method for assessing the cognitive representation of medication. *Psychol Health* 1999; 14: 1–24.
23. Chua SS, Lai PS, Tan CH, et al. The development and validation of the Malaysian Medication Adherence Scale (MALMAS) among patients with 2 type diabetes in Malaysia. *Int J Pharm Pharm Sci* 2013; 5: 790–794.
24. Griffith S. A review of the factors associated with patient compliance and the taking of prescribed medicines. *Br J Gener Pract* 1990; 40: 114–116.
25. Kapoor S and Narayanan A. Leakage and the reproducibility crisis in ML-based science. arXiv preprint arXiv:2207.07048. 2022. https://doi.org/10.48550/arXiv.2207.07048.
26. Kuhn M and Johnson K. *Applied predictive modeling*, Vol. 26. Springer, 2013.
27. Nti IK, Nyarko-Boateng O and Aning J. Performance of machine learning algorithms with different K values in K-fold CrossValidation. *Int J Inf Technol Comput Sci* 2021; 13: 61–71.
28. Vapnik V, Guyon I and Hastie T. Support vector machines. *Mach Learn* 1995; 20: 273–297.
29. Gama J. A linear-Bayes classifier. In: Ibero-American Conference on Artificial Intelligence, Atibaia, Brazil, 19–22 November 2000, pp. 269–279. Berlin/Heidelberg: Springer.
30. Cessie SL and Houwelingen JV. Ridge estimators in logistic regression. *J R Stat Soc Ser C Appl Stat* 1992; 41: 191–201.
31. Saunders C, Gammerman A and Vovk V. Ridge regression learning algorithm in dual variables. In: ICML'98: Proceedings of the Fifteenth International Conference on Machine Learning, San Francisco, CA, USA, 24–27 July 1998.
32. Breiman L. *Classification and regression trees*. Routledge, 2017.
33. Breiman L. Bagging predictors. *Mach Learn* 1996; 24: 123–140.
34. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal* 2002; 38: 367–378.
35. Fung G, Mangasarian O and Shavlik J. Knowledge-based support vector machine classifiers. *Adv Neural Inf Process Syst* 2002; 15: 537–544.
36. Freund Y and Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997; 55: 119–139.
37. Cover T and Hart P. Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 1967; 13: 21–27.
38. Song YY and Ying LU. Decision tree methods: applications for classification and prediction. *Shanghai Arch Psychiatry* 2015; 27: 130.
39. McCallum A and Nigam K. A comparison of event models for naive Bayes text classification. In: AAAI'98 Workshop on Learning for Text Categorization, Madison, WI, USA, 26–27 July 1998, Vol. 752, pp. 41–48.
40. Zhu H, Williams CKI, Rohwer R. Gaussian regression and optimal finite dimensional linear models. 1997. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b2d5fff7600abe81da7b1e5af749ed9511867aed.
41. Gu Y, Zalkikar A, Liu M, et al. Predicting medication adherence using ensemble learning and deep learning models with large scale healthcare data. *Sci Rep* 2021; 11: 18961.
42. Wu XW, Yang HB, Yuan R, et al. Predictive models of medication non-adherence risks of patients with T2D based on multiple machine learning algorithms. *BMJ Open Diabetes Res Care* 2020; 8: e001055.
43. Daghistani T and Alshammari R. Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. *J Adv Inf Technol* 2020; 11: 78–83.
44. Karanasiou GS, Tripoliti EE, Papadopoulos TG, et al. Predicting adherence of patients with HF through machine learning techniques. *Healthc Technol Lett* 2016; 3: 165–170.
45. Delpino FM, Costa AK, Farias SR, et al. Machine learning for predicting chronic diseases: a systematic review. *Public Health* 2022; 205: 14–25.
46. Kohavi R and John GH. Wrappers for feature subset selection. *Artif Intell* 1997; 97: 273–324.
47. Yoon S, Kwan YH, Yap WL, et al. Factors influencing medication adherence in multi-ethnic Asian patients with chronic diseases in Singapore: a qualitative study. *Front Pharmacol* 2023; 14: 1124297.
48. Levy R. Medication use by ethnic and racial groups: policy implications. *J Pharm Health Serv Res* 2010; 1: 15–22.
49. Chawa MS, Yeh HH, Gautam M, et al. The impact of socioeconomic status, race/ethnicity, and patient perceptions on medication adherence in depression treatment. *Prim Care Companion CNS Disord* 2020; 22: 26869.
50. Tanna S, Ogwu J and Lawson G. Hyphenated mass spectrometry techniques for assessing medication adherence: advantages, challenges, clinical applications and future perspectives. *Clin Chem Lab Med* 2020; 58: 643–663.
51. Lam WY and Fresco P. Medication adherence measures: an overview. *BioMed Res Int* 2015; 2015: 217047.
52. Fisher RJ, Rawal S, Hochstein B, et al. Development and validation of the SDR-O: a new measure of socially desirable responding in organizations. *Pers Individ Dif* 2024; 222: 112597.
53. Moon SJ, Lee WY, Hwang JS, et al. Correction: accuracy of a screening tool for medication adherence: a systematic review and meta-analysis of the Morisky Medication Adherence Scale-8. *PLoS One* 2018; 13: e0196138.
54. Aziz F, Malek S, Ali AM, et al. Determining hypertensive patients' beliefs towards medication and associations with medication adherence using machine learning methods. *PeerJ* 2020; 8: e8286.

55. Son YJ, Kim HG, Kim EH, et al. Application of support vector machine for prediction of medication adherence in heart failure patients. *Healthc Inform Res* 2010; 16: 253–259.
56. Mirzadeh SI, Arefeen A, Ardo J, et al. Use of machine learning to predict medication adherence in individuals at risk for atherosclerotic cardiovascular disease. *Smart Health* 2022; 26: 100328.
57. Li J, Cheng K, Wang S, et al. Feature selection: a data perspective. *ACM Comput Surv* 2017; 50: 1–45.
58. Vomlel J, Kruzık H, Tuma P, et al. Machine learning methods for mortality prediction in patients with st elevation myocardial infarction. *Proceedings of WUPES* 2012; 2012: 204–213.
59. Huyard C, Haak H, Derijks L, et al. When patients' invisible work becomes visible: non-adherence and the routine task of pill-taking. *Sociol Health Illn* 2019; 41: 5–19.
60. Dang L, Ananthasubramaniam A and Mezuk B. Spotlight on the challenges of depression following retirement and opportunities for interventions. *Clin Interv Aging* 2022: 1037–1056.
61. Christopher CM, Blebil AQ, Bhuvan KC, et al. Medication use problems and factors affecting older adults in primary healthcare. *Res Soc Adm Pharm* 2023; 19: 1520–1530.
62. Kvarnström K, Airaksinen M and Liira H. Barriers and facilitators to medication adherence: a qualitative study with general practitioners. *BMJ Open* 2018; 8: e015332.
63. Pantaleon L. Why measuring outcomes is important in health care. *J Vet Intern Med* 2019; 33: 356–362.
64. Fernandez-Lazaro CI, et al. Medication adherence and barriers among low-income, uninsured patients with multiple chronic conditions. *Res Soc Adm Pharm* 2019; 15: 744–753.
65. Norris P, Tordoff J, McIntosh B, et al. Impact of prescription charges on people living in poverty: a qualitative study. *Res Soc AdmPharm* 2016; 12: 893–902.
66. Aldeer M, Alaziz M, Ortiz J, et al. A sensing-based framework for medication compliance monitoring. In: Proceedings of the 1st ACM International Workshop on Device-Free Human Sensing, New York, NY, USA, 10 November 2019, pp. 52–56.
67. Di Palo KE, Patel K and Kish T. Risk reduction to disease management: clinical pharmacists as cardiovascular care providers. *Curr Probl Cardiol* 2019; 44: 276–293.
68. Gupta L, Khandelwal D, Lal PR, et al. Factors determining the success of therapeutic lifestyle interventions in diabetes – role of partner and family support. *Eur Endocrinol* 2019; 15: 18.
69. Matossian C. Noncompliance with prescribed eyedrop regimens among patients undergoing cataract surgery – prevalence, consequences, and solutions. *US Ophthal Rev*. 2020; 13: 18–22.
70. Farrell B, French Merkley V and Ingar N. Reducing pill burden and helping with medication awareness to improve adherence. *Can Pharm/Rev Pharmaciens Canada* 2013; 146: 262–269.
71. Usha P and Anuradha MP. An evaluation of feature selection methods performance for dataset construction. In: Subhashini N, Ezra MAG and Liaw SK (eds) Futuristic Communication and Network Technologies: Select Proceedings of VICFCNT 2021, Volume 1. Singapore: Springer Nature Singapore, 2023, pp. 115–128.
72. Lin LS. Improving small sample prediction performance via novel nonlinear interpolation virtual sample generation with self-supervised learning. *Inf Sci* 2024; 678: 121044.
73. Jin J, Yin F, Xu Y, et al. Learning a model with the most generality for small-sample problems. In: Proceedings of the 2022 5th International Conference on Algorithms, Computing and Artificial Intelligence, Sanya, China, 23–25 December 2022, pp.1–6.
74. Wang X, Zhang L and He T. Learning performance prediction-based personalized feedback in online learning via machine learning. *Sustainability* 2022; 14: 7654.
75. Peticca B, Prudencio TM, Robinson SG, et al. Challenges with non-descriptive compliance labeling of end-stage renal disease patients in accessibility for renal transplantation. *World J Nephrol* 2024; 13: 1–5.