Analysis of Machine Learning Models Predicting Quality of Life for Cancer Patients

Miloš Savić
Vladimir Kurbalija
Mihailo Ilić
Mirjana Ivanović
Dušan Jakovetić
Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad
Novi Sad, Serbia
{svc,kurba,milic,mira,dusan.jakovetic}@dmi.uns.ac.rs

Antonios Valachis
Department of Oncology, Faculty of Medicine and Health, Örebro University
SE 70182 Örebro, Sweden
antonios.valachis@oru.se

Serge Autexier
Johannes Rust
German Research Center for Artificial Intelligence GmbH, Cyber-Physical Systems
Bremen, Germany
{serge.autexier,johannes.rust}@dfki.de

Thanos Kosmidis
Care Across Ltd
London, England
thanos.kosmidis@careacross.com

ABSTRACT

Quality of life (QoL) is one of the major issues for cancer patients. With the advent of medical databases containing large amounts of relevant QoL information it becomes possible to train predictive QoL models by machine learning (ML) techniques. However, the training of predictive QoL models poses several challenges mostly due to data privacy concerns and missing values in patient data. In this paper, we analyze several classification and regression ML models predicting QoL indicators for breast and prostate cancer patients. Two different approaches are employed for imputing missing values. The examined ML models are trained on datasets formed from two databases containing a large number of anonymized medical records of cancer patients from Sweden. Two learning scenarios are considered: centralized and federated learning. In the centralized learning scenario all patient data coming from different data sources is collected at a central location prior to model training. On the other hand, federated learning enables collective training of machine learning models without data sharing. The results of our experimental evaluation show that the predictive power of federated models is comparable to that of centrally trained models for short-term QoL predictions, whereas for long-term periods centralized models provide more accurate QoL predictions.

MEDES '21, November 1–3, 2021, Virtual Event, Tunisia

ACM ISBN 978-1-4503-8314-1/21/11…$15.00
https://doi.org/10.1145/3444757.3485103

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Applied computing** → **Health informatics**.

KEYWORDS

Quality of Life, Cancer Patients, Predictive Models, Federated Learning, Breast Cancer, Prostate Cancer

ACM Reference Format:
Miloš Savić, Vladimir Kurbalija, Mihailo Ilić, Mirjana Ivanović, Dušan Jakovetić, Antonios Valachis, Serge Autexier, Johannes Rust, and Thanos Kosmidis. 2021. Analysis of Machine Learning Models Predicting Quality of Life for Cancer Patients. In *International Conference on Management of Digital EcoSystems (MEDES '21), November 1–3, 2021, Virtual Event, Tunisia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3444757.3485103

1 INTRODUCTION

Nowadays people of all ages suffer from different chronic diseases. The one of important trends in medical domains is to find out adequate services and support for such patients. Information and Communication Technologies offers powerful tools, devices and services for collection of wide range of patient's data: digital health records, wearable devices, dashboards, unobtrusive sensors and other types of smart devices. Using advanced and effective techniques of Artificial intelligence (AI) and Machine learning (ML) [13] such data are processed and can be used to improve patient's health status and different quality of life (QoL) aspects [12, 14].

Predictive analytics in medicine and healthcare plays essential role and recently is more concentrated on obtaining personalized interventions, treatment and improvement of QoL issues. QoL issues depend on type of disease and patient's status (beginning disease phase, after surgery, during treatment, follow-ups, etc.) and in this paper we will concentrate on QoL issues for breast and prostate

cancer, like: Depression, Anxiety, Influence of working environment, Sleep quality, Influence on physical activities, Fatigue, etc. General approach for assessment of patient's QoL issues is based on results of different questionnaires. Patient Reported Outcome Measures (PROMs) and Patient Reported Experience Measures (PREMs) are widely used high-quality questionnaires for checking patients' perceptions about their health and about experiences after a treatment/intervention.

Research results presented in the paper are achieved within ASCAPE project: Artificial intelligence supporting cancer patients across Europe (https://ascape-project.eu/) where two the most prevalent types of cancer are considered: breast and prostate cancer. One of the main purposes of the project is to employ powerful AI/ML mechanisms to support cancer patients' health status and QoL.

Within ASCAPE different validated questionnaires will be used to capture the particular QoL issues depending on type of cancer. Such questionnaires are good instruments to minimize the risk for measurement bias, and to enable the better ASCAPE ability for suitable predictions and interventions. Clinical ASCAPE partners identified 15 QoL issues for breast and 12 for prostate cancer that will be predicted through AI-based models.

To the best of our knowledge ASCAPE is a unique research project that prospectively investigate an AI-based approach, towards a personalised follow-up strategy for cancer patients focusing on their QoL issues. The aim of the project is to offer substantial benefits for after-treatment health-related QoL improvements.

Importance of applying adequate approaches to properly process huge amounts of sensitive medical data is evident. Two approaches to train predictive models are prevalent: centralized and distributed. In the centralized approach patients' data collected from different sources are stored locally in clinics and model training is performed locally as well. However, recently popular decentralized ML technique is Federated learning (FL) which enables the quality use and learning from decentralized data. Patients' data collected by clinics could be adequately processed by applying FL approach as it enables the training of shared global models with a central server, while keeping all the sensitive data in local. The primary purpose of FL in ASCAPE is to enable democratized access to ML models promoting patient QoL without revealing private or sensitive patient data.

Presented experimental results are based on already existing retrospective datasets of an ASCAPE clinical partner. The aim of conducted experiments was to identify appropriate ML model for breast and prostate cancer patients. However, numerous ML techniques which are utilized during project showed to be very promising and we will continue our research activities on prospectively collected data expecting accurate and reliable behavior of developed models.

The rest of paper is organized as follows. The second Section is devoted to related work. The central, third Section, is focused on application of ML techniques for predictive QoL models. Section four is devoted to the description of experimental datasets. Experimental results are discussed in Section five. The last section brings concluding remarks.

2 RELATED WORK

The tracking and monitoring of QoL parameters has attracted many research attention recently. This has a particular importance in chronic diseases for the prevention and early detection of symptoms and signs. The proper monitoring should provide the positive effect on patient's quality of life, economic impact and resource management. As expected, substantial amounts of data on quality of life are combined into clinical trials using a variety of instruments [15].

Regular assessment of QoL parameters of chronic patients has an impact on physician-patient communication and result in benefits for some patients, who had a better QoL [17]. Furthermore, continuous collection of medical attributes and growth of the database could allow the selection of proper medical variables and the selection of adequate models for a more accurate prediction of QoL [15].

Several recent studies emphasized the importance of applying AI and data mining (DM) techniques in successful prediction of QoL issues. For example, in [16] authors used QoL issues to monitor patients' clinical condition and health status for the patients with ultra-rare autosomal recessive disease Alkaptonuria. One of the main problems in the follow-up of patients with ultra rare diseases is the lack of a standardized methodology to assess disease severity or response to treatment and QoL scores could be a successful way to monitor such patients. Also, different machine learning approaches (Linear Regression, Neural networks and k-nearest-neighbor) were implemented with the aim to perform a prediction of QoL scores based on clinical data.

The authors in [12] present the basis of the DIAL system which represents an early warning system for the early detection of a deteriorating QOL score in the hemodialysis population using machine learning algorithms. Here two models (classification tree and Naïve Bayes) were generated to predict an increase or decrease of 5% in a patient's QoL score over one month. The classification tree was selected as the better model with an area under curve (AUC) of 83.3% and accuracy of 81.9%. The authors concluded that their system DIAL, if implemented on a larger scale, is expected to help patients in terms of ensuring a better QoL and a reduction in the financial burden in the long term.

The predictor of quality of life called the Better Life Index (BLI) was used for QoL assessment in [6]. It is based on the measurement of different aspects of human life in the whole population: environment, jobs, health, civic engagement, governance, education, access to services, housing, community, and income. The paper presents a supervised machine-learning analytical model that predicts the life satisfaction score using several DM models like: decision tree, elastic net, neural network, random forest, support vector machine, etc. The results showed that the ensemble model based on the stacked generalization framework is a significantly better predictor of the life satisfaction of a nation, compared to base models.

QoL are studied in different cancer related diseases since they present a common chronic conditions nowadays. For example, in [3] authors presented the project which tries to design a new patient reported outcome measure to assess QoL issues for patients with locally recurrent rectal cancer. The authors identified the fact that it is very tedious to administer simultaneously several questionnaires required for QoL assessment especially in patients with major limitations caused by the disease. So it is essential to find alternatives which largely replace questionnaires, and their approach involves use of biometric devices with the help of DM techniques.

Some successful applications of QoL prediction in cancer diseases also includes: the comparison of the pre and the post treatment quality of life in cervical cancer patients [8], construction of machine learning and statistical models for prediction of gastrourinary symptoms and QoL issues following prostate radiation treatment [20], and the prediction of 5-year lung cancer survival on the basis of QoL issues [14].

To the best of our knowledge, this study is the first one which tries to develop a QoL prediction model for breast and prostate cancer patients. The study [20] also investigates prostate cancer patients but focus only on one type of intervention - radiation treatment. Here we will try to encompass both types of cancer patients (breast and prostate) with as many as possible common attributes, with all types of common interventions, and with a set of relevant QoL attributes.

3 MACHINE LEARNING TECHNIQUES FOR PREDICTIVE QOL MODELS

The QoL indicator of a cancer patient can be predicted either by a classification or a regression machine learning model depending on the type of the indicator. In this work we consider two types of QoL indicators: (1) binary indicators indicating whether the patient will experience QoL related symptoms after diagnosis (e.g., anxiety and depression) and (2) numeric indicators indicating overall QoL of the patient reported by filling an appropriate questionnaire (e.g., the QoL score of the LISAT-11 questionnaire [10]). Binary classification models are the most appropriate for the first type of QoL indicators, while for the second type of QoL indicators predictions can be obtained by regression models.

For predictive QoL models (as well as for medical predictive models in general) it is important to distinguish two practical scenarios: centralized and decentralized (federated). In the centralized scenario all anonymized training data coming from different medical organizations is collected to a central place where machine learning models are trained. However, in majority of cases medical organizations are not willing to share anonymized patient data, or it is not even allowed for them to do so according to governmental regulations and laws. One solution in this case is to apply federated learning [18] of predictive QoL models in which several edge nodes collectively train predictive QoL models by exchanging locally updated machine learning models instead of sharing training datasets. Edge nodes are computational devices owned by medical organizations and deployed within their computational infrastructure, so training datasets never leave the boundaries of data owners.

3.1 Centralized QoL Predictive Models

The following machine learning algorithms are examined for training centralized binary classification models predicting binary QoL indicators: NB (Naive Bayes), KNN (K nearest neighbors), SVM (support vector machines), DT (decision trees) and RF (random forests).

NB is a probabilistic classification algorithm learning a predictive model giving the most probable class (positive or negative in the case of binary classification) for a given data instance (in our case a patient described by a set of features). Class probabilities are computed using conditional probability estimates learned from a training dataset under the assumption that features describing data instances are conditionally independent.

KNN is a lazy learning algorithm. The class for a given data instance is predicted by majority voting from the classes of the K closest data instances belonging the training dataset according to some distance functions (e.g., the Euclidean or Manhattan distance).

SVM classifiers are based on the idea of using linear models to identify non-linear boundaries of classes. This is achieved by transforming data instances into a new higher dimensional space using a non-linear mapping. Quadratic programming algorithms are then employed in the higher dimensional space to determine the maximum margin hyperplane separating instances from different classes.

Decision-tree based classifiers make predictions according to decision trees constructed from the training dataset by a recursive divide-and-conquer algorithm utilizing some information theoretic measure (e.g, information gain or the Gini impurity). This means that decision trees are formed in the divisive manner from the root of the tree to its leafs. The underlying information theoretic measure is used to find and select the most discriminative feature to form a corresponding node in the tree and split training data to recursively form its subtrees.

A random forest is an ensemble of decision trees learned from bootstrapped samples of the training dataset. The RF algorithm employs a feature bagging procedure to determine a random subset of features for learning individual decision trees. The class for a given input data instance is then determined as the most frequent class predicted considering all decision trees in the ensemble.

For predicting numeric QoL indicators we examine the following algorithms for learning regression models: LINEAR (linear regression), RIDGE (ridge regression), LASSO (lasso regression), ELASTICN (elastic net regression), KRIDGE (kernel ridge regression), SVM (regression by support vector machines), RF (regression by random forests), and KNN (K-nearest neighbours regression).

The linear regression algorithms determine coefficients of a linear model by minimizing the residual sum of squares (RSS) between real values of the target variable and predictions derived from the model. Ridge, Lasso and Elastic Net find linear models by minimizing RSS with incorporated regularization penalties: Ridge incorporates the L2 regularization penalty, Lasso is based on the L1 regularization penalty, while Elastic Net uses both previously mentioned penalties. Kernel Ridge regression performs Ridge regression in a space obtained by a non-linear mapping of the training dataset. SVM, RF and KNN are adaptations of the corresponding classification algorithms for regression tasks.

In our experiments we have used the Scikit-learn machine learning library [11] to develop a set of Python modules for training centralized QoL predictive models.

3.2 Federated QoL Predictive Models

A federated model is a machine learning model collectively trained by several edge nodes running federated learning clients. Each federated learning client has its own dataset for training the model and those local training datasets are never exchanged among federated learning clients participating in federated learning. The federated learning process is coordinated by a federated learning server. The

main purpose of the federated learning server is to enable the exchange of locally updated federated models among federated learning clients.

Two basic federated learning schemas are incremental and concurrent. Let C_1, C_2, .., C_k denote k federated learning clients each having its own training dataset D_i ($i \in [1 .. k]$). In the incremental federated learning scheme federated learning clients incrementally build a machine learning model from the first to the last client. This means that C_1 creates M on D_1 and sends it to the federated learning server. Then, C_2 retrieves M from the federated learning server, updates it on D_2 and returns the updated model back to the server. Each next federated learning client does exactly the same until the last client C_k.

In the concurrent federated learning scheme M is collectively trained in parallel. In the first step each federated learning client C_i creates its own model M_i on D_i. All models are then sent to the federated learning server which averages M_1 to M_k into a single model M. Once all edge nodes have submitted their local models, the global model is updated using the Federated Averaging [5] approach. The federated learning server then sends M to all federated learning clients which update M on their local datasets and the updated models are returned back to the federated learning server for the second averaging. The previous operation is repeated for an arbitrary number of learning rounds and the averaged model after the last learning round is the final federated model.

Neural networks are the most natural model choice for federated learning for the following two reasons: (1) neural networks can be incrementally updated, and (2) neural networks can be easily averaged by averaging edge weights and biases. Since we deal with two types of predictive problems we also have two types of federated neural networks:

(1) Federated neural networks for regression. The last layer of such neural networks contains exactly one node activated by the linear function. In our work we use the mean squared error (MSE) as the loss function when training regression neural networks.
(2) Federated neural networks for binary classification. In this case, one node contained in the last layer is activated by the sigmoid function. Output values higher than 0.5 indicate the positive class, while values lower than 0.5 correspond to the negative class. The binary cross-entropy function is used at the loss function when optimizing parameters of binary classification neural networks.

Nodes in hidden layers of both types of federated neural networks are activated by the ReLU activation function. We also consider two mechanisms to prevent overfitting: dropout and regularization strategies (the kernel, bias and activation regularization).

To compare federated QoL models to centralized QoL models we have developed a federated learning simulator based on the Tensorflow machine learning library [1]. The realized simulator supports both incremental and concurrent federated learning mode for an arbitrary number of simulated edge nodes (federated learning clients). The architecture of a federated neural network can be specified by providing its type (regression or binary classification), the number of hidden layers and the number of nodes per hidden layer. The user can also specify the number of epoch (learning rounds) and the batch size (the number of training instances propagated through the network when updating model parameters). At the beginning, the simulator divides training data (training folds when the k-fold cross-validation is applied to evaluate models) into p stratified parts, where p is the number of simulated edge nodes. Then, simulated edge nodes use their part of training data when creating or updating Tensorflow neural networks by the Adam optimization algorithm [7]. In the case of concurrent federated learning, the averaging of Tensorflow neural networks formed by simulated edge nodes is performed after each epoch. More precisely, we utilized the callback mechanism provided by Tensorflow to implement the federated client-server communication. A socket connection is open between an edge node and the coordinating server once a client wishes to update a global model, and at key training steps such as the end of each epoch, the edge node sends it's model updates to the server.

3.3 Missing Value Inference (MVI)

Datasets for training machine learning models often contain missing values. This is also the case with the experimental datasets used in this work. To train machine learning models it is necessary to infer and fill missing values of predictor features or, alternatively, to remove data instances containing missing values. For the missing value inference (MVI) we use two methods provided by the Scikit-learn library: (1) simple MVI and (2) iterative MVI. The simple MVI fills missing values using a simple approach: all missing values for a feature f are filled with the mean of existing values in f.

The iterative MVI is based on the idea to train a regression model for each feature containing missing values [2]. The regression model for f is trained based on values of other predictor features. Then, missing values for f are filled based on predictions of its regression model. After obtaining predictions for all missing values, the iterative MVI repeats the whole procedure for predefined number of times in a round-robin fashion (i.e., predicted missing values of f together with known values are then used to retrain the regression model for other features also containing missing values). The predictions of the final round are then used to fill missing values.

3.4 Model Evaluation

To estimate errors of the examined regression models we use the mean absolute error (MAE):

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \quad (1)$$

where n is the number of data instances in a test dataset (or a test fold), y_i the real value of the target variable for i-th instance and $\hat{y}_i$ is the predicted value of the target value according to a regression model.

In our experimental evaluation of QoL regression-based predictive models, MAE estimates of the examined regression models are also compared to MAE estimates of the so-called DUMMY regression algorithm. The DUMMY regression algorithm always predict the same value for a target feature: the mean of the target feature from a training dataset (or training folds).

Binary classifiers predicting QoL issues are evaluated using accuracy, precision, recall and F_1 scores. The accuracy of a binary

classifier is equal to the number of correctly classified instances divided by the total number of instances in a test dataset (or a test fold). Precision and recall metrics are defined per class. The precision for a class c (c is the positive or negative class) is the number of instances correctly classified to c divided by the total number of instances classified to c. On the other hand, the recall for c is the number of instances correctly classified to c divided by the total number of instances that belong to c. Precision and recall score for the whole binary classifier are obtained by averaging precision and recall scores per class. Since precision and recall measure two different aspects of classifier's performance it is useful to aggregate them into a single score. The usual way to aggregate precision and recall is to compute the F_1 score is which is the harmonic mean of precision (P) and recall (R):

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}. \quad (2)$$

All above-mentioned model evaluation metrics are estimated by the 10-fold cross validation procedure. In the 10-fold cross-validation, an input dataset is split into 10 folds. Then, 10 iterations of training and evaluation are performed. In the i-th iteration, i-th fold is used to compute evaluation metrics of models trained on the remaining folds. In this way we obtain 10 estimates of evaluation metrics that are then averaged into final estimates.

The effects of missing value inference to the performance of predictive QoL models is investigated by comparing models trained on datasets to which the simple MVI is applied and models trained on datasets obtained after the iterative MVI.

4 EXPERIMENTAL DATASETS

Previously described machine learning models are evaluated on datasets formed from two databases: ORB and BcBase. BcBase is a population-based research database containing data about early breast cancer patients from three healthcare regions in Sweden (accounting for nearly 60% of the total population) in terms of patient and tumor characteristics, treatment strategies and prescribed medications. The database does not include QoL indicators captured through questionnaires, but from prescribed medications it is possible to derive the presence of certain QoL issues. From BcBase we formed 4 datasets for training and evaluating binary classifiers deciding whether a patient will suffer from anxiety (BcBase-Anxiety), depression (BcBase-Depression), insomnia (BcBase-Insomnia) and pain (BcBase-Pain) after breast cancer treatment. All BcBase datasets contain 18988 data instances (patients) described by 97 predictor features.

ORB is a database containing data about patients with localised prostate cancer treated with radiotherapy in the Örebro healthcare region. It includes data on patient and tumour characteristics, treatment approaches, dosimetric parameters regarding radiotherapy, side effects based on direct questions or validated questionnaires (IPSS, IIEF-5) and QoL issues based on a validated questionnaire (LISAT-11). ORB contains 2466 health records with follow-up examinations repeated at six months intervals and LISAT-11 QoL scores at the time of the diagnosis and three different times relative to the date of diagnosis at months 36, 60 and 120. From ORB data we created six datasets with the naming scheme ORB-n-m for training regression models predicting the LISAT-11 QoL score at month m considering all patient data collected up to month n. The created datasets are: ORB-30-36, ORB-30-60, ORB-30-120, ORB-54-60, ORB-54-120 and ORB-108-120. For example, ORB-30-36 is used to train regression models predicting the LISAT-11 QoL score in month 36 considering data collected up to month 30 as predictor variables. The number of instances (patients) in ORB datasets ranges from 1138 in ORB-30-36 to 610 in ORB-108-120, while the number of predictor variables ranges from 96 in ORB-30-36 to 158 in ORB-108-120 (the number of patients decreases due to recoveries, dropouts and deaths, while the number of predictors increases due to longer time intervals. Also some patients got their diagnosis less than 10 years ago. They haven't had their follow-up yet).

5 RESULTS AND DISCUSSION

In this section we first present the evaluation of centralized machine learning models trained on experimental datasets described in previous section. The influence of missing value inference to the performance of those models is then discussed. Finally, we examine federated QoL machine learning models and compare them to their centralized counterparts.

5.1 Evaluation of Centralized QoL Models

The performance of binary classification models trained on BcBase datasets obtained by the 10-fold cross-validation are summarized in Tables 1, 2, 3, and 4. All classifiers were trained and evaluated on BcBase datasets obtained after missing value inference by the iterative MVI method.

Table 1: Evaluation of binary classification models on BcBase-Anxiety

Classifier	Accuracy	Precision	Recall	F_1
RF	0.673	0.484	0.535	0.514
SVM	0.698	0.411	0.349	0.5
NB	0.629	0.552	0.553	0.552
KNN	0.682	0.458	0.529	0.507
DT	0.583	0.511	0.511	0.511

Table 2: Evaluation of binary classification models on BcBase-Depression

Classifier	Accuracy	Precision	Recall	F_1
RF	0.677	0.473	0.524	0.509
SVM	0.702	0.413	0.351	0.5
NB	0.566	0.534	0.543	0.551
KNN	0.688	0.462	0.536	0.509
DT	0.589	0.515	0.515	0.515

The SVM classifier exhibits the highest accuracy on three out of four BcBase datasets (BcBase-Anxiety, BcBase-Depression and BcBase-Pain). However, precision and recall scores of SVM on those three BcBase datasets are significantly lower compared to other

Table 3: Evaluation of binary classification models on BcBase-Insomnia

Classifier	Accuracy	Precision	Recall	F_1
RF	0.538	0.526	0.535	0.532
SVM	0.541	0.533	0.539	0.537
NB	0.555	0.554	0.555	0.554
KNN	0.521	0.502	0.516	0.514
DT	0.516	0.515	0.515	0.515

Table 4: Evaluation of binary classification models on BcBase-Pain

Classifier	Accuracy	Precision	Recall	F_1
RF	0.698	0.486	0.554	0.518
SVM	0.714	0.417	0.357	0.5
NB	0.53	0.517	0.553	0.564
KNN	0.699	0.457	0.528	0.506
DT	0.604	0.522	0.522	0.522

classifiers. A more detailed examination of precision and recall per class revealed that SVM has zero precision and zero recall for the positive class (the presence of anxiety, depression and pain after treatment) and that it dominantly predicts the negative class (no negative QoL related symptoms). The highest accuracy of SVM on those three dataset is a consequence of its bias towards the negative class on class imbalanced datasets in which approximately 70% of the patients belong to the negative class and 30% to the positive class. Therefore, it can be concluded that accuracy is not an appropriate measure for comparing binary classification models trained on BcBase datasets. F_1 score is more adequate measure since it takes into account precision and recall of both classes. It can be observed that SVM exhibits the lowest F_1 score on those datasets where it has the highest accuracy. KNN has the lowest F_1 score on BcBase-Insomnia and the second lowest on other BcBase datasets. Thus, it can be concluded that those two methods are the worst performing binary classification models for BcBase datasets.

The largest F_1 score on three BcBase datasets (anxiety, depression and insomnia) is achieved by the NB classifier. The best model for the fourth dataset (pain) is DT, but its F_1 score is very close to the F_1 score of NB. Consequently, it can be concluded that NB is the best choice to train centralized QoL predictive models for BcBase datasets.

The results of the evaluation of centrally trained regression models on the ORB datasets are shown in Table 5 including the DUMMY regressor as the baseline. The best model (the lowest MAE) for ORB-30-36 is RF. For the rest of ORB datasets, the best performing model is LASSO. KNN is the worst performing regression algorithm on all ORB datasets: predictions made by this model are even more erroneous than predictions made by DUMMY. Excluding KNN, all others considered models exhibit smaller prediction errors compared to DUMMY except in one case: DUMMY is better than linear regression on ORB-30-120. The prediction errors of the best performing model are in the range [4.84, 6.47], which is an acceptable level of prediction errors taking into account that the target variable (the LISAT QoL index) is in the range [11, 66].

The improvement of LASSO (the best performing regression model) over DUMMY are significant for short term QoL predictions (30-36, 54-60, 108-120) when the reduction of the MAE score ranges from 20% to 30%. For medium term QoL predictions (30-60, 54-120) the improvements are between 10% and 15%. As expected, the lowest improvement is for long term QoL predictions on ORB-30-120 where the reduction of MAE scores is slightly higher than 5%.

Table 5: MAE scores of regression models on ORB datasets (best value per column is bolded)

Regressor	30-36	30-60	30-120	54-60	54-120	108-120
DUMMY	6.541	6.89	6.909	6.89	6.909	6.909
LINEAR	5.311	6.129	7.003	5.238	6.899	6.524
RIDGE	5.1	5.925	6.652	5.07	6.356	5.977
LASSO	5.089	**5.886**	**6.478**	**4.84**	**6.18**	**5.437**
ELASTICN	5.126	5.913	6.504	4.859	6.216	5.448
KRIDGE	5.147	5.958	6.75	5.115	6.492	6.155
SVM	6.519	6.773	6.871	6.772	6.859	6.875
RF	**5.051**	6.015	6.685	5.009	6.357	5.635
KNN	6.72	6.968	7.133	6.906	7.128	7.033

5.2 Influence of Missing Value Inference

For the best classification model (NB) and the best regression model (LASSO) we have examined the influence of two missing value inference methods (simple and iterative) to the models' performance. Figure 1 shows F_1 scored of NB for iterative and simple MVI methods. It can be seen that differences in F_1 scores are almost absent: the largest difference is equal to 0.004, which is less than 0.8% of the better score. Similar results are obtained for MAE scores of LASSO (Figure 2): the largest difference in MAE scores is equal to 0.12 (2.5% of the better score). Consequently, it can be concluded that the choice of the MVI method does not make a significant influence to the performance of examined models and that the simple MVI can be as equally effective as the iterative MVI.

5.3 Evaluation of Federated QoL Models

In experiments with simulated federated models, we have used different neural network architectures for BcBase and ORB datasets. A preliminary investigation, in which we have varied the number of hidden neural network layers between 1 and 10 and the batch size in the set {16, 64, 128, 256, 512}, showed that shallow neural networks (a small number of hidden layers) trained with a large batch size are more suitable for BcBase datasets, while deeper neural networks (a larger number of hidden layers) trained with a small batch size result with better predictive models for ORB datasets. We have simulated from 2 to 4 edge nodes training models in both incremental and concurrent federated learning mode. Performance metrics for simulated federated models were obtained in the same

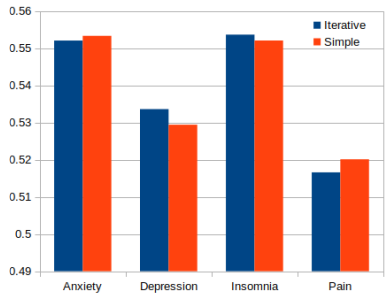

Figure 1: F_1 scores of NB on BC-Base datasets for two MVI methods (iterative and simple)

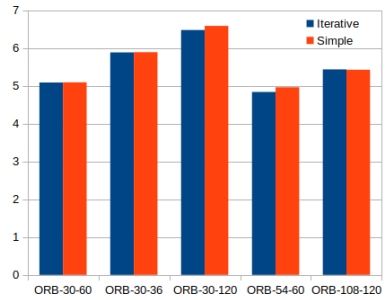

Figure 2: MAE scores of LASSO on ORB datasets for two MVI methods (iterative and simple)

way as for centrally-trained predictive models (by the 10-fold cross validation procedure).

The architecture of neural networks for federated binary classification models on the BcBase datasets consists of 4 hidden layers each having 20 nodes. Federated neural networks were trained in 200 epochs per simulated edge node with batch size equal to 512. The comparison of F_1 scores of centrally trained and simulated federated binary classification models is presented in Table 6. As baselines we use a centrally trained TensorFlow-based neural network (TFNN), the best centrally trained non-neural network classifier (Best C) and the worst centrally trained non-neural network classifier (Worst C). INC-k and CON-k denote federated neural network binary classification models trained in the incremental (INC) and concurrent (CON) learning mode for k simulated edge nodes.

For BcBase-Anxiety, Depression and Insomnia datasets, we have observed that federated models are significantly better than the worst performing local model (SVM and KNN depending on the dataset). F_1 scores of federated models are close to F_1 scores of NB which is the best performing centrally trained model. Federated models trained on BcBase-Pain have higher F_1 scores compared to the best performing centrally trained model on that dataset (DT). It is also important to emphasize that there are no significant differences in F_1 scores of incremental models and concurrent federated models. Additionally, the performance of federated models does not tend to significantly drop with the number of simulated edge nodes.

Table 6: F_1 scores of federated binary classification models on BcBase datasets.

	Anxiety	Depression	Insomnia	Pain
INC-2	0.536	0.512	0.546	0.542
INC-3	0.542	0.507	0.529	0.542
INC-4	0.539	0.515	0.538	0.532
CON-2	0.522	0.504	0.542	0.548
CON-3	0.512	0.519	0.55	0.534
CON-4	0.53	0.509	0.542	0.545
Best C	0.552	0.534	0.554	0.522
Worst C	0.411	0.413	0.502	0.457
TFNN	0.438	0.53	0.54	0.542

For federated regression models trained on the ORB dataset, we have used neural networks with 10 hidden layers each with 40 neurons. The training was performed in 200 epochs per simulated edge node with the batch size equal to 32. The obtained MAE scores are summarized in Table 7. For all six datasets, the best local model (LASSO) has lower prediction errors than simulated federated models. Large differences between federated models trained in different federated learning modes are absent. In contrast to federated models trained on BcBase datasets, here we can observe a tendency of increasing errors with the number of simulated edge nodes. Federated models are better than the DUMMY baseline for ORB-30-36, ORB-30-60, ORB-54-60 and ORB-108-120, but worse than DUMMY for ORB-30-120 and ORB-54-120. This result implies that different neural network architectures should be employed for short term and long term QoL predictions. Therefore, our subsequent work will be to examine a wider range of neural network architectures for federated regression and determine architectures providing satisfactory results long-term QoL predictions.

Table 7: MAE scores of federated regression models on ORB datasets.

	30-36	30-60	30-120	54-60	54-120	108-120
INC-2	6.012	6.775	7.488	5.931	7.188	6.625
INC-3	6.472	6.751	7.226	5.867	7.169	6.43
INC-4	6.595	7.042	7.463	6.22	7.206	6.484
CON-2	5.881	6.705	7.444	5.904	7.193	6.463
CON-3	6.404	6.826	7.427	5.986	7.098	6.327
CON-4	6.534	6.883	7.538	6.269	7.221	6.652
DUMMY	6.541	6.89	6.909	6.89	6.909	6.909
LASSO	5.089	5.886	6.478	4.84	6.18	5.437
TFNN	5.783	6.572	7.323	5.811	7.206	6.562

6 CONCLUSIONS AND FUTURE WORK

In this paper we have examined several classification and regression machine learning algorithms for training models predicting binary and numeric QoL indicators, respectively, for breast and prostate

cancer patients. The focus of our experimental evaluation was on two types of predictive models: (1) centrally-trained QoL models that are relevant either for individual data owners or for multiple data owners when it is allowed to collect training data in a central location, and (2) federated QoL models trained in distributed environments encompassing multiple data owners without data sharing. We also examined the influence on two different MVI algorithms on the performance of examined predictive models.

Our experimental evaluation on real datasets showed that numeric QoL indicators can be accurately predicted by both centrally-trained and federated models for short term future periods. Centrally-trained regression models provide also accurate long term predictions. On the other hand, federated regression models exhibit prediction errors close to the dummy regression model indicating that different neural network architectures should be employed for learning regression-based federated models providing short term and long term QoL predictions.

For classification-based models predicting binary QoL indicator it was observed that centrally-trained and federated models have comparable prediction performances. However, for both types of model we noticed that they achieve relatively low precision and recall scores for the minority class due to imbalanced training datasets indicating that appropriate data sampling techniques should be examined to form more class-balanced training datasets prior to model training. Our analysis of simple and iterative MVI algorithms showed that the choice of the MVI does not make a significant influence to the performance of examined models.

In our future work, we will also examine various feature selection techniques to identify the most relevant features for making QoL predictions and examine the performance of predictive QoL models trained on selected features. Having in mind that predictive QoL models should be robust against adversarial inference requests, we will also examine the influence of differential privacy mechanisms [4] to the accuracy of predictive QoL models. Furthermore, it would be interesting to examine federated learning approaches which are not based on neural networks like: decision trees [9] or logistic regression [19].

ACKNOWLEDGMENTS

This research was supported by the ASCAPE project. The ASCAPE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 875351.

REFERENCES

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A System for Large-Scale Machine Learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation* (Savannah, GA, USA) *(OSDI'16)*. USENIX Association, USA, 265–283.

[2] Stef Buuren and Catharina Groothuis-Oudshoorn. 2011. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45 (12 2011). https://doi.org/10.18637/jss.v045.i03

[3] Joaquim Gonçalves, Brígida Mónica Faria, Luís Paulo Reis, Victor Carvalho, and Álvaro Rocha. 2015. Data mining and electronic devices applied to quality of life related to health data. In *2015 10th Iberian Conference on Information Systems and Technologies (CISTI)*. 1–4. https://doi.org/10.1109/CISTI.2015.7170627

[4] Maoguo Gong, Yu Xie, Ke Pan, Kaiyuan Feng, and A.K. Qin. 2020. A Survey on Differentially Private Machine Learning [Review Article]. *IEEE Computational Intelligence Magazine* 15, 2 (2020), 49–64. https://doi.org/10.1109/MCI.2020.2976185

[5] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. 2021. Advances and Open Problems in Federated Learning. arXiv:1912.04977 [cs.LG]

[6] Maninder Kaur, Meghna Dhalaria, Pradip Sharma, and Jong Park. 2019. Supervised Machine-Learning Predictive Analytics for National Quality of Life Scoring. *Applied Sciences* 9 (04 2019), 1613. https://doi.org/10.3390/app9081613

[7] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[8] Satwant Kumar, Madhu Rana, Khushboo Verma, Narayanjeet Singh, Anil Sharma, Arun Maria, Gobind Singh, Harkiran Khaira, and Sunil Saini. 2014. PrediQt-Cx: Post Treatment Health Related Quality of Life Prediction Model for Cervical Cancer Patients. *PloS one* 9 (02 2014), e89851. https://doi.org/10.1371/journal.pone.0089851

[9] Qinbin Li, Zeyi Wen, and Bingsheng He. 2019. Practical Federated Gradient Boosting Decision Trees. arXiv:1911.04206 [cs.LG]

[10] Roland Melin, Kerstin Fugl-Meyer, and Axel Fugl-Meyer. 2003. Life satisfaction in 18-to 64-year-old Swedes: In relation to education, employment situation, health and physical activity. *Journal of rehabilitation medicine : official journal of the UEMS European Board of Physical and Rehabilitation Medicine* 35 (04 2003), 84–90. https://doi.org/10.1080/16501970306119

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[12] Shoab Saadat, Ayesha Aziz, Hira Ahmad, Hira Imtiaz, Zara Sohail, Alvina Kazmi, Sanaa Aslam, Naveen Naqvi, and Sidra Saadat. 2017. Predicting Quality of Life Changes in Hemodialysis Patients Using Machine Learning: Generation of an Early Warning System. *Cureus* 9 (09 2017). https://doi.org/10.7759/cureus.1713

[13] Jenni Sidey-Gibbons and Chris Sidey-Gibbons. 2019. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology* 19 (03 2019). https://doi.org/10.1186/s12874-019-0681-4

[14] Jinah Sim, Young Kim, Ju Kim, Jong Lee, Moon Soo Kim, Young Shim, Jae Zo, and Young Ho Yun. 2020. The major effects of health-related quality of life on 5-year survival prediction among lung cancer survivors: applications of machine learning. *Scientific Reports* 10 (07 2020), 10693. https://doi.org/10.1038/s41598-020-67604-3

[15] Richa Sinha and Wim Heuvel. 2011. A systematic literature review of quality of life in lower limb amputees. *Disability and rehabilitation* 33 (06 2011), 883–99. https://doi.org/10.3109/09638288.2010.514646

[16] Ottavia Spiga, Vittoria Cicaloni, Cosimo Fiorini, Alfonso Trezza, Anna Visibelli, Lia Millucci, Giulia Bernardini, Andrea Bernini, Barbara Marzocchi, Daniela Braconi, Filippo Prischi, and Annalisa Santucci. 2020. Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet Journal of Rare Diseases* 15 (12 2020). https://doi.org/10.1186/s13023-020-1305-0

[17] Galina Velikova, Laura Booth, Adam Smith, Paul Brown, Pamela Lynch, Julia Brown, and Peter Selby. 2004. Measuring Quality of Life in Routine Oncology Practice Improves Communication and Patient Well-Being: A Randomized Controlled Trial. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* 22 (03 2004), 714–24. https://doi.org/10.1200/JCO.2004.06.078

[18] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* 10, 2, Article 12 (Jan. 2019), 19 pages. https://doi.org/10.1145/3298981

[19] Shengwen Yang, Bing Ren, Xuhui Zhou, and Liping Liu. 2019. Parallel Distributed Logistic Regression for Vertical Federated Learning without Third-Party Coordinator. arXiv:1911.09824 [cs.LG]

[20] Zhijian Yang, Daniel Olszewski, Chujun He, Giulia Pintea, Jun Lian, Tom Chou, Ronald C. Chen, and Blerta Shtylla. 2021. Machine learning and statistical prediction of patient quality-of-life after prostate radiation therapy. *Computers in Biology and Medicine* 129 (2021), 104127. https://doi.org/10.1016/j.compbiomed.2020.104127