Article

Not peer-reviewed version

Optimizing Telehealth-Based Cancer Pain Management Through Machine Learning

Sergio Coluccia , Anna Crispo , Alessandro Ottaiano , Mariachiara Santorsola , Massimo Antonio Innamorato , Valentina Cerrone , Rosario De Feo , Dalila Esposito , Maria Pia Bruno , Francesco Sabbatino , Marco Cascella *

Posted Date: 21 October 2025

doi: 10.20944/preprints202510.1661.v1

Keywords: cancer pain; telemedicine; telehealth; machine learning; artificial intelligence

 doi:10.20944/preprints202510.1661.v1

Article

Optimizing Telehealth-Based Cancer Pain Management Through Machine Learning

Sergio Coluccia [1], Anna Crispo [2], Alessandro Ottaiano [2], Mariachiara Santorsola [2], Massimo Antonio Innamorato [3], Valentina Cerrone [4], Rosario De Feo [4], Dalila Esposito [4], Maria Pia Bruno [4], Francesco Sabbatino [4] and Marco Cascella [4,*]

[1] Branch of Medical Statistics, Biometry and Epidemiology "G. A. Maccacaro", Department of Clinical Sciences and Community Health, Dipartimento di Eccellenza 2023–2027, Università degli Studi di Milano, 20133 Milan, Italy

[2] Istituto Nazionale Tumori-IRCCS "Fondazione G. Pascale", via M. Semmola 9, 80131 Naples, Italy

[3] Pain Unit, Department of Neuroscience, Santa Maria delle Croci Hospital, AUSL Romagna, 48121 Ravenna, Italy

[4] Department of Medicine, Surgery and Dentistry "Scuola Medica Salernitana", University of Salerno, Salerno, Italy

* Correspondence: mcascella@unisa.it

Abstract

Background. Although telehealth strategies can be effectively used for managing cancer pain, identifying the best care pathway for tailoring interventions and allocating resources remains difficult. Artificial intelligence and machine learning (ML) may help clinicians develop more accurate strategies for predicting whether patients need remote consultations or in-person evaluations. **Methods.** Data from two cohorts of cancer pain patients were analyzed. Variables included sociodemographic and clinical data including age, sex, ECOG performance status, metastases, bone metastases, pain type, breakthrough cancer pain (BTCP), and rapid onset opioids (ROOs) therapy. The main outcome was the number of televisits (one versus multiple). After harmonizing the dataset, categorical variables were one-hot encoded, and age was standardized. Six models were tested: logistic regression, random forest (RF), gradient boosting machine (GBM), support vector machine (SVM), k-nearest neighbors (KNN), and multilayer perceptron (MLP). Training and tuning used a 7-repeated 5-fold cross-validation approach. Performance was evaluated on a hold-out test set using F1-score, accuracy, and AUC-ROC. A sensitivity analysis with two scenarios was performed to verify the effects of class weighting and excluding the cohort variable. **Results**. The final dataset included 270 patients. No variable was significantly linked to the number of televisits. F1-scores across models ranged from 0.33 (RF) to 0.65 (MLP), accuracy from 0.45 (RF) to 0.55 (SVM), and AUC-ROC from 0.43 (RF) to 0.65 (LR). DeLong tests showed no significant differences between algorithms ($p > 0.05$). While the MLP achieved the highest F1-score, it showed instability with 91% null F1-scores. Incorporating class weights slightly improved SVM (F1 = 0.58; AUC = 0.62) and LR (F1 = 0.53; AUC = 0.63), though not significantly. Removing the cohort variable reduced training time by about one hour and yielded similar results. **Conclusion**. Although no model demonstrated strong predictive power, this ML-based framework shows the potential of using structured telemedicine data to model clinical workload and optimize follow-up strategies in cancer pain care. Further studies with feature-rich datasets are needed to improve clinical usefulness.

Keywords: cancer pain; telemedicine; telehealth; machine learning; artificial intelligence

1. Introduction

Cancer pain is one of the most common and debilitating symptoms across all types of cancer, significantly reducing patients' quality of life [1]. Moreover, managing cancer-related pain can often

 doi:10.20944/preprints202510.1661.v1

be challenging [2]. For example, many patients with advanced disease struggle to attend frequent clinic visits due to their limited ability to perform daily activities or because of unmanageable pain and other symptoms [3]. In this complex situation, telehealth interventions can help bridge this gap by improving access to care and maintaining ongoing pain management [4]. Notably, recent evidence shows that telemedicine-based programs can lead to a reduction in pain intensity and pain interference with daily life compared to usual in-person care [5]. These findings highlight the feasibility of remote care options and the effectiveness and acceptability of these strategies as key components of cancer pain management.

However, implementing and optimizing telemedicine-based pain care pathways pose several challenges that must be addressed. Given the need for a personalized care process, cancer pain management often requires multidisciplinary input and flexible adjustment of therapy [6]. Therefore, scheduling a care pathway that includes both clinic visits and remote assessments is often complex. Key issues include identifying suitable assessment tools for remote pain evaluation, implementing reliable symptom monitoring devices, and determining when additional interventions or in-person exams are necessary [7].

Therefore, there is an urgent need to establish calibrated, dynamic protocols that strike the right balance between telehealth and in-person visits. These protocols are necessary for timely clinical assessments, additional diagnostic tests, or procedures. In the absence of comprehensive guidelines or large-scale trials in this field, developing such a structured telemedicine pathway depends on early experiences and expert consensus. While a hybrid care model that combines remote visits with on-site hospital care was demonstrated to be an effective and safe strategy for managing cancer pain, precise refinement of this process is essential [8]. It should incorporate patient feedback and outcomes [9]. For example, a telehealth-based pain management program in Italy combined virtual consultations with traditional clinic visits, resulting in positive outcomes regarding patient adherence and pain control. This initial experience demonstrated that a hybrid telemedicine approach can deliver high-quality care while reaching patients who live up to 500 km away from the cancer center [8–10].

On these premises, artificial intelligence (AI) and machine learning (ML) can be used to analyze multiple patient-related variables, such as demographic, clinical, and therapeutic metadata, to predict which patients will require more intensive follow-up and, ultimately, identify those likely to need additional remote consultations or earlier in-person visits [11].

2. Materials and Methods

2.1. Study Population and Ethics

The study population included adult patients treated for cancer pain at the Istituto Nazionale Tumori, Fondazione Pascale, Napoli, Italy, and at the Ruggi University Hospital, Salerno, Italy.

The local Medical Ethics Committees approved this study (protocol code 41/20 Oss; date of approval, 26 November 2020; Ethical Committee Campania 2, N°2024/28590, 3 April 2025), and all patients provided written informed consent. The investigation was conducted in accordance with the Declaration of Helsinki.

2.2. Datasets

Two cohorts of cancer patients, cohort 1 recorded during 2021 and cohort 2 in 2023/2025, were considered in the analysis. A single operator (Marco Cascella) performed the televisits for both cohorts. The same technology and processes (platform, devices, privacy policy) were adopted.

The date of any single televisit was recorded. The included variables were age, sex, tumor site, Eastern Cooperative Oncology Group (ECOG) performance status, presence of metastasis, bone metastasis, pain (breakthrough cancer pain, BTCP, no BTCP), type of pain (nociceptive, neuropathic/mixed), and rapid opioid onset (ROO) therapy (no, yes). The time-varying variables, such as ECOG, BTCP, pain type, and ROO therapy, were considered at the first televisit, according

 doi:10.20944/preprints202510.1661.v1

to the former dataset. Continuous measurements were shown as mean (standard deviation, SD) and median (interquartile range, IQR), while categorical measures were synthesized with absolute frequencies (percentages). Repositories are available at [12,13].

2.3. Dataset Harmonization and Preprocessing

This initial step involved adapting two datasets. For cohort 2, time-varying information about variables was collected and had to be aligned with the format of the cohort 1 dataset. Specifically, after transposing data into a wide format, we only considered information on age, sex, ECOG, (bone) metastasis, BTCP pain, pain type, and ROO therapy at the first televisit. Data on tumor sites and the presence of metastasis were also available, but they were excluded from the analysis due to the small number of observations and their strong correlation with other features.

The final data included 280 cancer patients from cohort 1 (n = 226) and cohort 2 (n = 54). Since 10 observations containing missing values were removed, a total of 270 complete observations were considered for the subsequent analysis.

The number of televisits was recorded in discrete values and categorized into one televisit or more televisits. One-hot encoding was applied to categorical variables, while age, the only continuous measure, was standardized based on the values in the training set.

2.4. ML Framework

2.4.1. ML Algorithms

Our analysis aimed to classify observations with one or more televisits. To showcase the full methods pipeline and ensure reproducibility, we outlined our framework following guidelines for the recommended use of supervised ML models [14,15]. Five ML algorithms and a multiple logistic regressor (LR) as a reference model were used for this analysis. Specifically, the random forest (RF) and the gradient boosting machine (GBM) are tree-based methods consisting of multiple decision trees (DTs), which individually process the input data and collectively produce the final output. RF is based on the bagging (bootstrap aggregating) method to build pseudo-uncorrelated trees that focus on different feature subsets as predictors for the outcome. This results in a stable algorithm capable of handling many covariates. An RF can be configured by hyperparameters, including the number of trees (ntree), the number of random features (mtry) per tree, and the minimum number of observations in a terminal node (node size). The boosting learner is an alternative approach applied to the sequential construction of DTs in the GBM ensemble method. After creating an initial DT, the prediction error is gradually reduced by subsequent trees. The growth of these trees is regulated by the number of trees, terminal node size, and the shrink (learning rate) parameter. The support vector machine (SVM) aims to find the best hyperplane that separates observations. This hyperplane maximizes the distance between the hyperplane and the nearest data points from both classes. The optimal hyperplane, known as a hard margin, maximizes these margins. When nonlinear relationships exist between the outcome and features, the probabilistic shape of the data is used to transform the input. We implemented kernel density using a radial basis function (RBF) kernel, a common approach in SVM classification tasks. Other hyperparameters include gamma, which relates to the RBF kernel, and the cost parameter (c or cost), which penalizes the model for misclassifications within the margins. Higher values enforce stricter separation. These parameters are usually tuned within known value ranges [16]. The k-nearest neighbors (KNN) algorithm is widely used for classification, where a new data point is classified based on the majority class among its k closest neighbors. The hyperparameter k must be specified before running the algorithm. The multilayer perceptron (MLP) is a type of artificial neural network composed of interconnected layers of neurons. Each neuron acts as a switch that transforms its input via an activation function and outputs a signal if a threshold is met. The neurons are organized in layers, with hidden layers in between the input and output layers. Each layer receives input from the previous layer, computes its output based on weights, and passes it forward. The final layer produces the network's output. Backpropagation involves iteratively passing the output backward through the network to adjust and optimize the

 doi:10.20944/preprints202510.1661.v1

weights, which control the strength of signal transmission. This enhances the network's predictive ability but also increases computational demands. We used the hyperbolic tangent as the activation function, as it proved more stable when handling outliers [17]. The network was built with one or two hidden layers (HLs). The first could vary from 1 to 5 neurons, and the second was set equal to the first. The learning rate determines how quickly weights are updated during training in a backpropagation process.

2.4.2. Data Splitting, Explorative Analysis, and ML Models Training and Selection

Data was split into a 75/25% training/test, balanced on the combination of the number of televisits and cohort. An explorative univariate nalysis consisted of detecting statistical differences across cohorts and showing crude associations between the number of televisits (One, More televisits) with the other variables. Tests included Pearson's chi-squared test for independence and Fisher's exact test. The Benjamini-Hochberg correction was applied to deal with multiple tests.

Five ML methods (RF, GBM, SVM, KNN, MLP) were implemented together with the reference LR model. The training was performed by an R-repeated K-fold cross-validation method. In particular, for each ML algorithm, we adopted R = 7 and K = 5: for a single combination of hyperparameters, the training set was split into 5 parts: 4 (n = 165) used for training each algorithm, and the last (n = 39) to validate the algorithm. A seed based on K and R was applied to allow the reproducibility of splitting. Observations were randomly given to the training or test set, balancing the proportion of the number of televisits together with the cohort. The algorithms characterized by the corresponding hyperparameters linked to the best performance were chosen.

Performances were calculated via the F1-score metric. The selected models were re-run on the whole train set (n = 204) and internally tested and compared with each other on the F1-score, Accuracy, and AUC-ROC measures calculated on the test set (n = 66). The Delong test [18] was adopted to detect pair-wise differences in terms of prediction performances, according to the AUC score.

A sensitivity analysis was conducted to identify potential changes in the performance of algorithms. Specifically, all ML pipelines were rerun under two scenarios: a) we applied class weights to the models based on the proportion of televisits and cohort membership; b) we removed the cohort variable from the feature set, resulting in an unbalanced dataset while keeping the same training and test set structure used in the main analysis. In scenario a), the function used to implement the MLP did not allow for direct input of class weights. Therefore, an oversampling procedure was designed to achieve a similar effect.

All tests and estimates were presented at the 95% confidence level. The analyses were implemented using the R software for statistical analysis, version 4.3.2, and the following packages were included: dplyr for data manipulation, gtsummary for data presentation, caret for the performance metrics computation, randomForest, gbm, e1071, class, neuralnet for algorithm implementation, purrr and pROC for AUC-ROC analysis, and ggplot to draw the figures. The workflow is available at [19].

3. Results

Demographic and clinical data were collected into two separate datasets. The two cohorts include 226 individuals from cohort 1 and 54 from cohort 2. The characteristics of the sample are described in Table 1. Mean age was 64.5 years (±12.4), half were males, the most prevalent tumor site was gastrointestinal (20.7%), followed by breast (12.9%). Most patients suffered from metastatic disease (90%), whereas half of them had bone metastasis. Almost 40% suffered from BTCP, while a similar prevalence was observed in neuropathic or mixed pain. Sixty percent of the sample underwent ROO therapy. The median number of televisits was 1 (IQR: 1, 3), with 143 patients having 1 televisit and 9% having more than four (see Table 2, in the middle). When analyzing the difference between cohorts (not shown in tables), cohort 2 had a significantly higher ECOG than cohort 1 (p =

 doi:10.20944/preprints202510.1661.v1

0.02), reported a higher prevalence of bone metastases (70.4% vs 45.1%, $p < 0.01$), and suffered from nociceptive pain more than cohort 1 (78.8 vs 59.1, $p = 0.01$).

Nevertheless, no differences in the number of televisits were observed by cohort ($p > 0.90$). When analyzing our main outcome (Table 1), no significant association were found between the number of televisits with the available set of features, even if an indication of relationship was visible for sex, where requiring more televisits was more prevalent among males than females (55.6% vs 44.8%, $p = 0.07$), and for type of pain: in particular patients undergoing more televisits had a more prevalent neuropathic or both pain compared to the one who underwent only one televisit ($p = 0.09$).

Table 1. Descriptive statistics and univariable analysis: overall sample distribution over all available features, by cohort, and by number of televisits.

	Overall	**Univariate analysis by Cohort**			**Univariate analysis by number of televisits**		
Characteristic	N = 280[1]	**Cohort 1**, N = 226[1]	**Cohort 2**, N = 54[1]	**p-value**[2]	**One**, N = 143[1]	**More**, N = 137[1]	**p-value**[2]
Number of televisits				0.423			
One	226 (80.7%)	111 (49.1%)	32 (59.3%)				
More	54 (19.3%)	115 (50.9%)	22 (40.7%)				
Number of televisits (numbers)[3]							
Mean (SD)	2.2 (2.1)	2.2 (1.9)	2.4 (2.8)				
Median (IQR)	1 (1, 3)	2 (1, 3)	1 (1, 3)				
Cohort							0.180
Cohort 1	226 (80.7%)				111 (77.6%)	115 (83.9%)	
Cohort 2	54 (19.3%)				32 (22.4%)	22 (16.1%)	
Age				0.333			0.726
Mean (SD)	64.5 (12.4)	64.1 (12.3)	66 (12.6)		66.5 (12.5)	62.4 (12)	
Median (IQR)	65 (57, 74)	65 (56, 74)	68.5 (58.3, 75.8)		68 (58, 76.0)	62.5 (54, 72)	
(Missing)	5	5	0		2	3	
Sex				0.226			0.073
Female	140 (50%)	109 (48.2%)	31 (57.4%)		79 (55.2%)	61 (44.5%)	

 doi:10.20944/preprints202510.1661.v1

Male	140 (50%)	117 (51.8%)	23 (42.6%)		64 (44.8%)	76 (55.5%)
ECOG (numbers)[3]						
Mean (SD)	2.5 (0.7)	2.4 (0.6)	3 (1)		2.5 (0.8)	2.5 (0.7)
Median (IQR)	2 (2, 3)	2 (2, 3)	3 (2, 4)		2 (2, 3)	2.0 (2, 3)
(Missing)	3	1	2		1	2
ECOG (class)			**0.019**			0.692
1-2	147 (53.1%)	127 (56.4%)	20 (38.5%)		77 (54.2%)	70 (51.9%)
3-4	130 (46.9%)	98 (43.6%)	32 (61.5%)		65 (45.8%)	65 (48.1%)
(Missing)	3	1	2		1	2
Tumor site[3]						
Bladder	18 (6.4%)	16 (7.1%)	2 (3.7%)		12 (8.4%)	6 (4.4%)
Breast	36 (12.9%)	30 (13.3%)	6 (11.1%)		13 (9.1%)	23 (16.8%)
Endocrine	20 (7.1%)	11 (4.9%)	9 (16.7%)		6 (4.2%)	14 (10.2%)
Gastrointestinal	58 (20.7%)	45 (19.9%)	13 (24.1%)		38 (26.6%)	20 (14.6%)
Gynecological	11 (3.9%)	10 (4.4%)	1 (1.9%)		5 (3.5%)	6 (4.4%)
Head/neck	14 (5%)	14 (6.2%)	0 (0.0%)		6 (4.2%)	8 (5.8%)
Kidney	6 (2.1%)	6 (2.7%)	0 (0.0%)		1 (0.7%)	5 (3.6%)
Lung	39 (13.9%)	31 (13.7%)	8 (14.8%)		22 (15.4%)	17 (12.4%)
Melanoma/skin	5 (1.8%)	5 (2.2%)	0 (0.0%)		3 (2.1%)	2 (1.5%)
Prostate	22 (7.9%)	16 (7.1%)	3 (5.6%)		11 (7.7%)	8 (5.8%)
Soft tissue/bones	32 (11.4%)	22 (9.7%)	0 (0.0%)		9 (6.3%)	13 (9.5%)
Others	19 (6.8%)	20 (8.8%)	12 (22.2%)		17 (11.9%)	15 (10.9%)
Metastasis				0.329		0.112

 doi:10.20944/preprints202510.1661.v1

No	31 (11.1%)	23 (10.2%)	8 (14.8%)		20 (14.0%)	11 (8.0%)	
Yes	249 (88.9%)	203 (89.8%)	46 (85.2%)		123 (86.0%)	126 (92.0%)	
Bone metastasis				**0.002**			0.901
No	142 (50.7%)	125 (55.3%)	17 (31.5%)		72 (50.3%)	70 (51.1%)	
Yes	138 (49.3%)	101 (44.7%)	37 (68.5%)		71 (49.7%)	67 (48.9%)	
BTCP				0.083			0.252
No	169 (60.4%)	142 (62.8%)	27 (50%)		91 (63.6%)	78 (56.9%)	
Yes	111 (39.6%)	84 (37.2%)	27 (50%)		52 (36.4%)	59 (43.1%)	
Type of pain				**0.011**			0.086
Nociceptive	175 (62.7%)	133 (59.1%)	42 (77.8%)		96 (67.6%)	79 (57.7%)	
Neuropathic/mixed	104 (37.3%)	92 (40.9%)	12 (22.2%)		46 (32.4%)	58 (42.3%)	
(Missing)	1	1	0		1	0	
ROO therapy				0.826			0.691
No	110 (39.4%)	88 (39.1%)	22 (40.7%)		58 (40.6%)	52 (38.2%)	
Yes	169 (60.6%)	137 (60.9%)	32 (59.3%)		85 (59.4%)	84 (61.8%)	
(Missing)	1	1	0		0	1	

[1]n (%); [2]Pearson's Chi-squared test, Wilcoxon rank sum test. The Benjamini & Hochberg correction for multiple testing; [3]not included in the multiple testing correction. Abbreviations: BTCP, breakthrough cancer pain; ROO, rapid onset opioid; ECOG, Eastern Cooperative Oncology Group.

Table 2. Hyperparameters to optimize the model on the 7-repeated 5-cross validation technique.

	Hyperparameters/characteristics	Execution time to train and select (mins)	R package used for implementation
RF	• ntree: 20 to 100, step of 10 • mtry: 3 to 6 • nodesize: 9, 11, 17	0.86	randomForest (v. 4.7.1.1)
GBM	• ntree: 20 to 100, step of 10 • node size: 9, 11, 17 • shrink: 0.001, 0.005, and from 0.01 to 0.1, step of 0.01	3.39	gbm (version 2.2.2)
SVM	• gamma: 2 raised to the power of: from -15 to 3, step of 2 • c: 2 raised to the power of: from -5 to 15, step of 2	3.60	e1071 (version 1.7.14)
KNN	k from 2 to 5	0.01	class (version 7.3.22)

 doi:10.20944/preprints202510.1661.v1

MLP	• learning rate: 0.001, 0.005, and from 0.01 to 0.19, step of 0.02 • HL1: from 1 to 5 • HL2: from 0 and < HL1	231.30	neuralnet (version 1.44.2)
LR	logit link		stats (from R version 4.3.2)

Abbreviations: RF, random forest; GBM, gradient boosting machine; SVM, support vector machine; KNN, k-nearest neighbors; MLP, multilayer perceptron; LR, logistic regression.

3.1. ML Modelling

Table 2 summarizes the hyperparameter grids and computational time required to optimize each ML algorithm using the 7-repeated 5-fold cross-validation procedure. For each model, a range of key hyperparameters was explored to identify the configuration providing the best validation performance.

The RF and GBM models were tuned by varying the number of trees (ntree = 20–100), node size (9–17), and the number of features randomly sampled at each split (mtry = 3–6 for RF) or the shrinkage rate (shrink = 0.001–0.1 for GBM). The SVM was optimized over exponential grids of gamma (2^−15 to 2^3) and cost (2^−5 to 2^15). For the KNN algorithm, k varied from 2 to 5. The MLP network was tuned through combinations of learning rates (0.001–0.19, step 0.02) and the number of neurons in the first (HL1 = 1–5) and second hidden layers (HL2 < HL1). The LR model used a logit link as a reference.

In terms of computational cost, ensemble, and kernel-based methods (GBM = 3.39 min; SVM = 3.60 min; RF = 0.86 min) showed moderate training times, while the neural network (MLP) required substantially longer optimization (≈ 231 min). Simpler algorithms, such as KNN and LR, trained almost instantaneously.

The optimal hyperparameters and linked F1-scores on the test set with accuracy and AUC-ROC scores are shown in **Table 3**. Each model was retrained using the best configuration obtained from the 7-repeated 5-fold cross-validation and then tested on unseen data (n = 66).

Among ensemble methods, the RF achieved an F1-score of 0.43 and an accuracy of 0.49 (95% CI: 0.36–0.61), while the GBM performed similarly (F1 = 0.42; AUC = 0.42). The SVM with $\gamma = 2$ and cost = 0.125 showed a moderate improvement (F1 = 0.55; accuracy = 0.55; AUC = 0.55). The KNN algorithm with k = 3 produced comparable results (F1 = 0.45; accuracy = 0.48; AUC = 0.45).

Notably, the MLP, optimized with a learning rate of 0.05 and two hidden layers (HL1 = 4, HL2 = 3), achieved the highest discrimination capacity (F1 = 0.65; AUC = 0.65), although with relatively lower accuracy (0.48) due to class imbalance. Finally, the LR model yielded intermediate performance (F1 = 0.47; accuracy = 0.52; AUC = 0.47), confirming its role as a stable but less flexible baseline classifier.

Despite the differences in mean values, no statistically significant difference among models was detected (DeLong test, $p > 0.05$), indicating comparable predictive capacities under the same data conditions.

Table 3. Confusion matrices and associated performance metrics on the test set (n = 66).

	Optimal hyperparameters	TN	FN	FP	TP	F1-score	Accuracy	AUC-ROC
RF	ntree: 70; mtry: 3; node size: 9	19	19	15	19	0.43	0.49 (95%CI: 0.36, 0.61)	0.43
GBM	ntree: 80; node size: 11; shrink: 0.02	17	19	17	13	0.42	0.46 (0.33, 0.67)	0.42
SVM	gamma: 2; c: 0.125	18	14	16	18	0.55	0.55 (0.42, 0.61)	0.55
KNN	k: 3	18	18	16	14	0.45	0.48 (0.36, 0.61)	0.45
MLP	learning rate: 0.05; HL1: 4; HL2: 3	0	0	34	32	0.65	0.48 (0.36, 0.61)	0.65

 doi:10.20944/preprints202510.1661.v1

LR	20	18	14	14	0.47	0.52 (0.39, 0.64)	0.47

Abbreviation: TP, true positive; FN, false negative; FP, false positive; TN, true negative; AUC-ROC, area under the curve – receiving operating characteristic curve; RF, random forest; GBM, gradient boosting machine; SVM, support vector machine; KNN, k-nearest neighbors; MLP, multilayer perceptron; LR, logistic regressor.

Comparisons across AUCs reported values from 0.43 (RF) to 0.65 (LR) (Figure 1). GBM was particularly unable to detect observations correctly, leading to an accuracy < 0.5, and to a statistically significantly worse AUC than SVM and LR (p-value from the Delong test = 0.04 and 0.03, respectively).

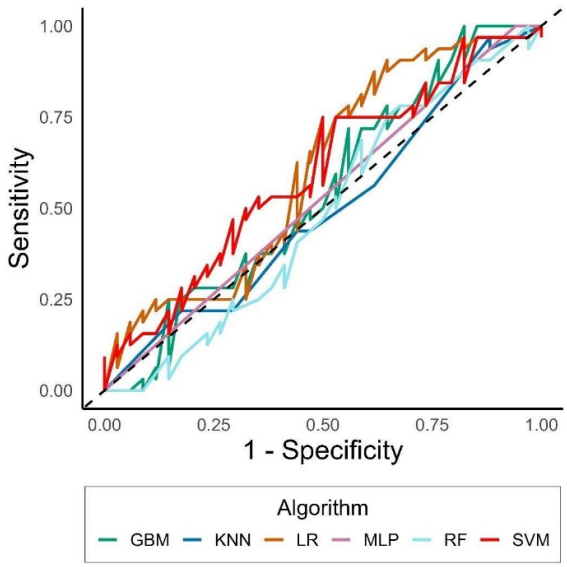

Figure 1. AUC-ROC curves calculated on the test set (n = 66) of the six models. No significant differences were found across algorithms from the main analysis.

The sensitivity analysis did not detect particular differences in the classification behaviors across models, even if an additional time was needed to train the MLP (6.8 hours), due to the oversampling routine that was applied when tackling the sub-analysis on the class weight, where they were given as follows: One televisit and cohort 1: n = 81, weight = 1.02; one televisit and cohort 2: n = 24, weight = 3.46; more televisits and cohort 1: n = 83, weight = 1.0; more televisits and cohort 2: n = 16, weight = 5.19. In contrast, the computational time was reduced by about one hour when the analysis was performed without considering the cohort among the features. Figure 2 reports the AUC-ROCs from these additional analyses.

 doi:10.20944/preprints202510.1661.v1

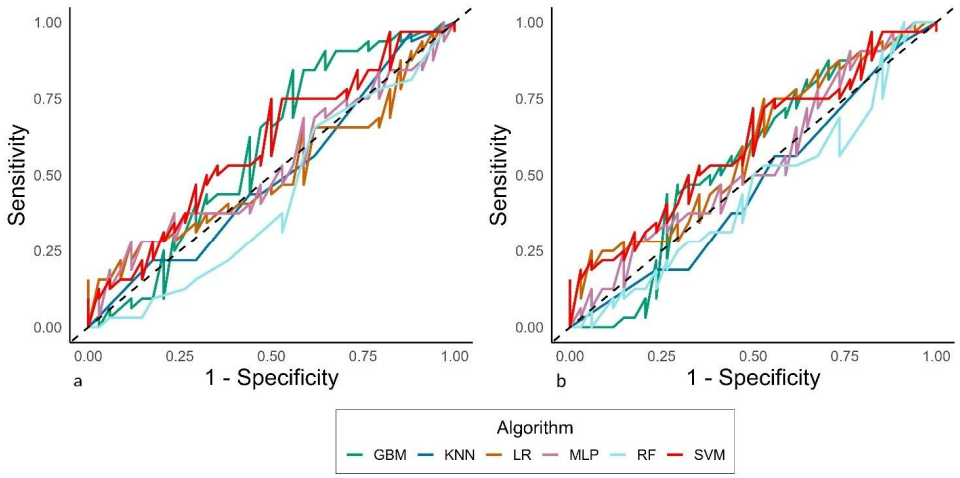

Figure 2. AUC-ROC curves calculated on the test set (n = 66) of the six models from the sensitivity analysis: a) adding the weights procedure to the models; b) removing the cohort among the features (see methods for the description). No significant differences were found across algorithms from the main analysis.

4. Discussion

Although telemedicine represents a suitable opportunity for care delivery, clinical practice indicates that scheduling and balancing in-person and telehealth-based consultations is often challenging [7,10,20]. In the complex context of cancer pain management, ML-driven predictions can enable clinicians to tailor care interventions and allocate resources.

In our previous ML investigation, we found that models identified specific factors that increased the likelihood of requiring multiple teleconsultations. These factors included younger patient age (<55 years), certain cancer types (e.g., lung cancer), and the occurrence of BTCP episodes. The analysis also suggested that elderly patients (>75 years) with advanced disease and bone metastases might benefit from closer telemedicine monitoring [21]. In this study, which focused on using ML models to predict the need for additional visits, despite implementing rigorous validation procedures, model performance remained modest across all tested algorithms. Probably, these results reflect intrinsic limitations of the dataset rather than methodological flaws.

A first critical issue relates to the risk of underfitting. In ML contexts, underfitting occurs when algorithms fail to learn sufficient patterns from the data, producing models that perform similarly to random classification [22]. This is a common problem in small and homogeneous datasets, especially when the outcome variable exhibits class imbalance or low variability [23]. Specifically, the dataset included a limited number of predictors, most of which were categorical and derived from baseline assessments. Consequently, the limited feature complexity may have affected the models' ability to capture nonlinear relationships and hidden interactions among variables. Nevertheless, the simplicity of the dataset and the small number of clinically meaningful predictors likely reduced the capacity of the algorithms to generalize. This is a key issue as cancer pain is influenced by multifactorial dynamics, including psychological, pharmacologic, and social variables, and probably they were not fully captured in this analysis [24]. Therefore, expanding future datasets to include pivotal information such as patient-reported outcomes, analgesic adjustments, as well as clinical input obtained from wearable sensor data and digital technology tools [25], could substantially enhance predictive modeling.

The presence of two distinct cohorts may have introduced heterogeneity that impaired model learning. Statistically significant differences were observed in ECOG performance status, pain type, and prevalence of bone metastases. These inter-cohort disparities may represent background differences not directly associated with the outcome (number of televisits), thereby acting as noise that limits pattern recognition. Similar findings have been described in ML studies using heterogeneous clinical populations, where variations in data source or clinical practice reduce

 doi:10.20944/preprints202510.1661.v1

algorithmic stability. They represent crucial gaps for AI deployment into clinical practice [26]. Andersen et al. [27], for instance, discussed how variations in clinical data sources and operational settings can compromise the stability of AI and ML models, thereby requiring continuous monitoring practices and model updating strategies. However, the inclusion of a comprehensive sensitivity analysis confirmed the robustness of the models across different training conditions. Techniques for handling class imbalance, such as class-weight adjustment or oversampling of minority classes, have been extensively documented in the literature to mitigate bias due to class or cohort imbalance [28–30]. In our analysis, given the application of class-weight adjustments and testing the exclusion of the cohort variable, potential sources of bias related to class imbalance and cohort heterogeneity were effectively mitigated. Therefore, model performance remained stable and was not driven by data distribution artifacts.

From a methodological perspective, other strengths could be highlighted. For example, the application of a repeated k-fold cross-validation scheme (7×5) represents a robust internal validation strategy, consistent with international guidelines such as the data, optimization, model and evaluation (DOME) approach [14], and the multidisciplinary recommendations for developing ML models in biomedicine [15]. Moreover, the integration of multiple algorithms (tree-based, kernel-based, and neural models) within the same pipeline provides a comparative view of how different learning paradigms behave in a low-dimensional telemedicine dataset.

Additionally, another potential strength is the translational applicability of the proposed framework. Given the availability of larger and more granular datasets, ML-based care processes can be directly applied to predict high-frequency teleconsultations, optimize clinician workload, and tailor follow-up [31]. According to key objectives in modern telehealth oncology practice, these tools could assist healthcare providers in proactively identifying patients who might benefit from closer remote monitoring or careful in-person evaluations [5]. Therefore, on these premises, future studies should aim to integrate temporal and textual features (e.g., free-text notes, symptom trajectory), to evaluate model calibration and explainability using techniques such as SHAP or LIME [32], and to perform external validation on independent cohorts. Addressing these aspects will increase generalizability and clinical trustworthiness, aligning with emerging AI quality standards and the European AI Act framework [33].

5. Conclusions

Despite limitations, this ML analysis confirms the feasibility of applying predictive analytics to telemedicine data for cancer pain management. Importantly, results highlight the importance of dataset quality, feature richness, and cohort homogeneity in developing reliable predictive models. Future research should expand data sources and exploit more sophisticated architectures to strengthen the integration of AI into personalized telehealth care.

Author Contributions: Conceptualization, S.C., M.C. and A.C.; methodology, S.C., M.S. and F.S.; software, M.A.I., D.E. and V.C.; validation, A.O., M.P.B. and R.D.F.; formal analysis, M.S., S.C. and A.O.; investigation, D.E., V.C. and A.C.; resources, R.D.F., M.P.B. and F.S.; data curation, M.A.I., D.E. and M.S.; writing—original draft preparation, S.C., A.O. and A.C.; writing—review and editing, M.C., M.P.B. and F.S.; visualization, V.C., R.D.F. and M.A.I.; supervision, M.C., F.S. and A.O.; project administration, M.C., A.C. and R.D.F.; funding acquisition, M.C., A.O. and F.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The local Medical Ethics Committees approved this study (protocol code 41/20 Oss; date of approval, 26 November 2020; Ethical Committee Campania 2, N°2024/28590, 3 April 2025), and all patients provided written informed consent.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

 doi:10.20944/preprints202510.1661.v1

Data Availability Statement: The data that support the findings of this study are available from the corresponding author, M.C., upon reasonable request. The data are not publicly available due to ethical restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Snijders RAH, Brom L, Theunissen M, van den Beuken-van Everdingen MHJ. Update on Prevalence of Pain in Patients with Cancer 2022: A Systematic Literature Review and Meta-Analysis. Cancers (Basel). 2023;15(3):591. doi: 10.3390/cancers15030591.
2. Carbonara L, Casale G, De Marinis MG, Bosetti C, Valle A, Carinci P, D'andrea MR, Corli O. Adherence to ESMO guidelines on cancer pain management and their applicability to specialist palliative care centers: An observational, prospective, and multicenter study. Pain Pract. 2024 Oct 3. doi: 10.1111/papr.13418.
3. van Roij J, Brom L, Youssef-El Soud M, van de Poll-Franse L, Raijmakers NJH. Social consequences of advanced cancer in patients and their informal caregivers: a qualitative study. Support Care Cancer. 2019;27(4):1187-1195. doi: 10.1007/s00520-018-4437-1.
4. Chen W, Huang J, Cui Z, Wang L, Dong L, Ying W, Zhang Y. The efficacy of telemedicine for pain management in patients with cancer: a systematic review and meta-analysis. Ther Adv Chronic Dis. 2023 Feb 17;14:20406223231153097. doi: 10.1177/20406223231153097.
5. Buonanno P, Marra A, Iacovazzo C, Franco M, De Simone S. Telemedicine in Cancer Pain Management: A Systematic Review and Meta-Analysis of Randomized Controlled Trials. Pain Med. 2023;24(3):226-233. doi: 10.1093/pm/pnac128.
6. Porzio G, Capela A, Giusti R, Lo Bianco F, Moro M, Ravoni G, Zułtak-Baczkowska K. Multidisciplinary approach, continuous care and opioid management in cancer pain: case series and review of the literature. Drugs Context. 2023;12:2022-11-7. doi: 10.7573/dic.2022-11-7.
7. Cascella M, Marinangeli F, Vittori A, Scala C, Piccinini M, Braga A, Miceli L, Vellucci R. Open Issues and Practical Suggestions for Telemedicine in Chronic Pain. Int J Environ Res Public Health. 2021;18(23):12416. doi: 10.3390/ijerph182312416.
8. Cascella M, Schiavo D, Grizzuti M, Romano MC, Coluccia S, Bimonte S, Cuomo A. Implementation of a Hybrid Care Model for Telemedicine-based Cancer Pain Management at the Cancer Center of Naples, Italy: A Cohort Study. In Vivo. 2023;37(1):385-392. doi: 10.21873/invivo.13090.
9. Cascella M, Coluccia S, Grizzuti M, Romano MC, Esposito G, Crispo A, Cuomo A. Satisfaction with Telemedicine for Cancer Pain Management: A Model of Care and Cross-Sectional Patient Satisfaction Study. Curr Oncol. 2022 Aug 4;29(8):5566-5578. doi: 10.3390/curroncol29080439.
10. Cascella M, Innamorato MA, Natoli S, Bellini V, Piazza O, Pedone R, Giarratano A, Marinangeli F, Miceli L, Bignami EG, Vittori A. Opportunities and barriers for telemedicine in pain management: insights from a SIAARTI survey among Italian pain physicians. J Anesth Analg Crit Care. 2024 Sep 17;4(1):64. doi: 10.1186/s44158-024-00202-1.
11. Li YH, Li YL, Wei MY, Li GY. Innovation and challenges of artificial intelligence technology in personalized healthcare. Sci Rep. 2024;14(1):18994. doi: 10.1038/s41598-024-70073-7.
12. Cascella M. PainDatafor_Telemedicine_ML [Data set] Zenodo. 2022 https://doi.org/10.5281/zenodo.6944310.
13. Cascella, M. (2025). Telemedicine_Cancer_Pain_RUGGI [Data set]. Zenodo. https://doi.org/10.5281/zenodo.17241215
14. Walsh I, Fishman D, Garcia-Gasulla D, Titma T, Pollastri G; ELIXIR Machine Learning Focus Group; Harrow J, Psomopoulos FE, Tosatto SCE. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021;18(10):1122-1127. doi: 10.1038/s41592-021-01205-4. Erratum in: Nat Methods. 2021;18(11):1409-1410. doi: 10.1038/s41592-021-01304-2.
15. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, Shilton A, Yearwood J, Dimitrova N, Ho TB, Venkatesh S, Berk M. Guidelines for Developing and Reporting Machine Learning Predictive Models in

 doi:10.20944/preprints202510.1661.v1

Biomedical Research: A Multidisciplinary View. J Med Internet Res. 2016;18(12):e323. doi: 10.2196/jmir.5870.

16. Hsu C, Chang C, Lin C. A Practical Guide to Support Vector Classification. 2003. Available from: https://api.semanticscholar.org/CorpusID:2443126
17. Cuevas FG, Schmid H. Robust Non-linear Normalization of Heterogeneous Feature Distributions with Adaptive Tanh-Estimators. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics. In: Proceedings of Machine Learning Research. 2024;238:406-414 Available from https://proceedings.mlr.press/v238/guimera-cuevas24a.html.
18. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44, 837–845.
19. Coluccia, S. (2025). Telemedicine_ML_Workflow. Zenodo. https://doi.org/10.5281/zenodo.17392667.
20. Wyte-Lake T, Cohen DJ, Williams S, Bailey SR. Balancing Access, Well-Being, and Collaboration When Considering Hybrid Care Delivery Models in Primary Care Practices with Team-Based Care. J Am Board Fam Med. 2025 Sep 15;38(3):475-489. doi: 10.3122/jabfm.2024.240388R2.
21. Cascella M, Coluccia S, Monaco F, Schiavo D, Nocerino D, Grizzuti M, Romano MC, Cuomo A. Different Machine Learning Approaches for Implementing Telehealth-Based Cancer Pain Management Strategies. J Clin Med. 2022;11(18):5484. doi: 10.3390/jcm11185484.
22. Nazha A, Elemento O, McWeeney S, Miles M, Haferlach T. How I read an article that uses machine learning methods. Blood Adv. 2023;7(16):4550-4554. doi: 10.1182/bloodadvances.2023010140.
23. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6. doi: 10.1186/s12864-019-6413-7.
24. Bennett MI, Kaasa S, Barke A, Korwisi B, Rief W, Treede RD; IASP Taskforce for the Classification of Chronic Pain. The IASP classification of chronic pain for ICD-11: chronic cancer-related pain. Pain. 2019 Jan;160(1):38-44. doi: 10.1097/j.pain.0000000000001363.
25. Goncalves Leite Rocco P, Reategui-Rivera CM, Finkelstein J. Telemedicine Applications for Cancer Rehabilitation: Scoping Review. JMIR Cancer. 2024;10:e56969. doi: 10.2196/56969.
26. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019;17(1):195. doi: 10.1186/s12916-019-1426-2.
27. Andersen ES, Birk-Korch JB, Hansen RS, Fly LH, Röttger R, Arcani DMC, Brasen CL, Brandslund I, Madsen JS. Monitoring performance of clinical artificial intelligence in health care: a scoping review. JBI Evid Synth. 2024;22(12):2423-2446. doi: 10.11124/JBIES-24-00042.
28. Chen W, Yang K, Yu Z. et al. A survey on imbalanced learning: latest research, applications and future directions. Artif Intell Rev. 2024;57:137. Doi: 10.1007/s10462-024-10759-6.
29. Gurcan F, Soylu A. Learning from Imbalanced Data: Integration of Advanced Resampling Techniques and Machine Learning Models for Enhanced Cancer Diagnosis and Prognosis. Cancers (Basel). 2024;16(19):3417. doi: 10.3390/cancers16193417.
30. Araf I, Idri A, Chairi I. Cost-sensitive learning for imbalanced medical data: a review. Artif Intell Rev. 2024;57: 80. doi: 10.1007/s10462-023-10652-8.
31. Christopoulou SC. Machine Learning Models and Technologies for Evidence-Based Telehealth and Smart Care: A Review. BioMedInformatics. 2024; 4(1):754-779. doi: 10.3390/biomedinformatics4010042.
32. Contreras J, Winterfeld A, Popp J, Bocklitz T. Spectral Zones-Based SHAP/LIME: Enhancing Interpretability in Spectral Deep Learning Models Through Grouped Feature Analysis. Anal Chem. 2024 Oct 1;96(39):15588-15597. doi: 10.1021/acs.analchem.4c02329.
33. Commission of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence (AI Act). Available from: https://artificialintelligenceact.eu/the-act/