Radiomics and Machine Learning for Skeletal Muscle Injury Recovery Prediction

Vasileios Eleftheriadis, José Raul Herance Camacho, Valentina Paneta, Bruno Paun, Carolina Aparicio, Vanesa Venegas, Mario Marotta, Marc Masa, George Loudos, and Panagiotis Papadimitroulas

***Abstract*—Radiomics as a novel quantitative approach to medical imaging is an emerging area in the field of radiology. Artificial intelligence offers promising tools for exploiting and analyzing radiomics. The objective of the present study is to propose a methodology for the design, development, and evaluation of machine learning (ML) models for the prediction of the recovery progress of skeletal muscle injury over time in rats using radiomics. Radiomics were extracted from contrast enhanced computed tomography (CT) data and ML algorithms were trained and compared for their predictive value based on different CT imaging parameters. Ten different ML regression algorithms were tested and the optimal combination of radiomics for each algorithm and CT imaging parameter settings combination was studied. The best ensemble learning model, trained on the 70 kVp, 100 mA imaging parameter dataset, achieved a mean absolute error score of 1.22. The results suggest that radiomics extracted from CT images can be used as input in ML regression algorithms to predict the volume of a skeletal muscle injury in rats. Moreover, the results show that CT imaging settings impact the predictive performance of the ML regression models, indicating that lower values of tube current and peak kilovoltage contribute to more accurate predictions.**

***Index Terms*—Computed tomography (CT), machine learning (ML), muscle injury, preclinical imaging, prediction model, radiomics, recovery.**

Manuscript received 16 May 2023; accepted 28 June 2023. Date of publication 4 July 2023; date of current version 3 November 2023. This work was supported by the European Union's Horizon Research and Innovation Program under Grant 761031. *(Corresponding author: Panagiotis Papadimitroulas.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Vall d'Hebron Animal Ethics Committee under Application No. 52/17, and performed in line with the Spanish (Real Decreto 53/2013) and European (2010/63/UE) legislation.

Vasileios Eleftheriadis, Valentina Paneta, George Loudos, and Panagiotis Papadimitroulas are with the R&D Department, Bioemission Technology Solutions, 15343 Athens, Greece (e-mail: vasilis.eleftheriadis@bioemtech.com; vpaneta@bioemtech.com; george@bioemtech.com; panpap@bioemtech.com).

José Raul Herance Camacho, Bruno Paun, and Carolina Aparicio are with the Medical Molecular Imaging Group, Vall d'Hebron Research Institute, CIBER-BBN, CIBBIM-Nanomedicine, ISCIII, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, 08035 Barcelona, Spain (e-mail: raul.herance@vhir.org; brunopaun@gmail.com; carolina.aparicio@vhir.org).

Vanesa Venegas and Mario Marotta are with the Health & Biomedicine Department, Leitat Technological Center, 08225 Barcelona, Spain, and also with the Bioengineering, Cell therapy and Surgery in Congenital Malformations Laboratory, Vall d'Hebron Research Institute, CIBBIM-Nanomedicine, Hospital Universitari Vall d'Hebron, Universitat Autònoma de Barcelona, 08035 Barcelona, Spain (e-mail: mmarotta@leitat.org).

Marc Masa is with the Health & Biomedicine Department, Leitat Technological Center, 08225 Barcelona, Spain (e-mail: mmasa@leitat.org).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TRPMS.2023.3291848.

Digital Object Identifier 10.1109/TRPMS.2023.3291848

I. Introduction

BIOMEDICAL imaging, since its invention, has been an essential tool for clinical decision, medical intervention, and research. Recent advancements in computer technology, such as the increase in computing power, the digitalization of medical imaging, and thus the increase in dataset sizes and their harmonization, offer new possibilities of utilizing medical images not just as pictures intended solely for visual interpretation but as a source of data as well [1], [2].

The concept of radiomics was initially introduced in 2012 by Lambin et al. [3] and is a process of extracting high-dimensional feature data from digital medical images to quantitatively describe attributes of Regions of Interest (ROIs). The field of radiomics is based on the hypothesis that biomedical images contain information imperceptible by the human eye that reflects underlying pathophysiology and that these relationships can be revealed via quantitative image analysis [2]. This way, by quantifying differences in image intensity, shape, or texture, the use of radiomics does not only enhance the existing data available to clinicians with additional information but also helps overcoming the subjective nature of medical image interpretation [4].

Radiomics extraction can lead to a significantly vast number of features from each medical image, that are usually combined with artificial intelligence (AI) methodologies and more precisely with machine learning (ML) algorithms [7]. Moreover, the ability of radiomics to quantify textural information from biomedical images allows ML algorithms to focus on 1-D arrays of numerical input features instead of 3-D images.

Radiomics have primarily been applied in oncology, however, the potential benefits of radiomics are not limited to this field. Several recent studies in oncology utilize radiomics features in combination with AI techniques for cancer diagnosis, prediction, and management in a variety of organs and systems [8], such as prostate [9], [10], lung [11], [12], [13], kidney [14], brain [15], [16], liver [17], [18], adrenal gland [19], [20], and pituitary gland [21], [22]. Radiomics studies not related to oncology have started to appear as well, including studies on detection of cardiovascular risk factors on cardiac structure and tissue [23], classification and prediction of mild cognitive impairment and Alzheimer's disease [24], identification of temporal lobe epilepsy [25], diagnosis, classification and prediction of oral diseases [26], phenotyping of cardiovascular disease [27], diagnosis, prediction and prognosis of stroke [28], and placental tissue characterization [29]. In sports medicine, magnetic resonance imaging (MRI) has been used extensively in studies focused on classification and assessment of muscle injuries. In such studies researchers attempted to predict the time to return to sport after injury, or return to play (RTP), of athletes based on MRI muscle injury grade classification without being able to be conclusive about its predictive value [31], [32], [33], [34].

Regarding skeletal muscle injury recovery studies, Paun et al. [30] focused on the applicability of in vivo computed tomography (CT) imaging to track skeletal muscle lesion recovery over time in rats. The least absolute residual (LAR) method available in MATLAB was used to train an exponential model that predicts the recovery of skeletal muscle injury over time in rats. Feeding their model with the volume of the initial injury (Day 0) and the post-injury time, in days, Paun et al. achieved a mean root-mean-square error (RMSE) of 6.8 in their predictions. In the current study, we are assessing the ability of radiomic features to predict the recovery of skeletal muscle injury over time on the same CT dataset of Paun et al. when used as input in ML algorithms.

To the best of our knowledge, there are no studies where radiomics features, and AI have been used to predict the healing process of skeletal muscle injuries. Consequently, the aim of this study was to introduce a methodology of creating a model that predicts the recovery progress of skeletal muscle injury in rats by applying ML techniques on radiomics features data and comparing the predictive quality of different CT imaging parameter settings.

II. Materials and Methods

A. Data Description—Dataset Preclinical

In this study, we used the preclinical imaging dataset of skeletal muscle injuries in rats of Paun et al. study [30]. The dataset consists of CT images of 23 Wistar male adult rats. For the acquisition of all CT images a Quantum FX micro-CT scanner (PerkinElmer, Hopkinton, MA, USA) was used. In preparation for the CT studies the rats were anaesthetized and immobilized. Skeletal traumatic muscle injuries were induced by a transverse biopsy procedure in the muscle-tendon junction level of the rats' left leg medial gastrocnemius muscle by an 18-gauge biopsy needle with a 0.84-mm inner diameter [35]. To help distinguish between injury and neighboring tissue, iopamidol was administered to the rats as a contrast agent.

In order to track injury recovery at different time points, two studies were conducted, the Single Post-Injury study, where injury is monitored only once after the injury and the longitudinal one, where injury is tracked for several days post injury. In the Single Post-injury study, 20 rats were sorted into five separate groups ($n = 4$ per group) according to the single follow-up day at 2, 4, 7, 10, or 14 days after injury, respectively, providing in total 40 CT instances (two instances for each mouse, one at the day of injury and one at each follow-up day post injury). In the longitudinal study, three rats were imaged at all five mentioned follow-up days, providing a total of 18 CT instances (six instances per mouse, including the one at the day of injury). Each instance consists of four CT images of different value combinations of the imaging parameters peak kilovoltage (kVp) and tube current (mA) as follows: 90 kVp_200 mA, 70 kVp_200 mA, 90 kVp_100 mA, and 70 kVp_100 mA. A schematic representation instances' dataset can be seen in Table A1 (supplementary material). More details about the methods utilized for the acquisition of the CT studies' data can be found in the study by Paun et al. [30].

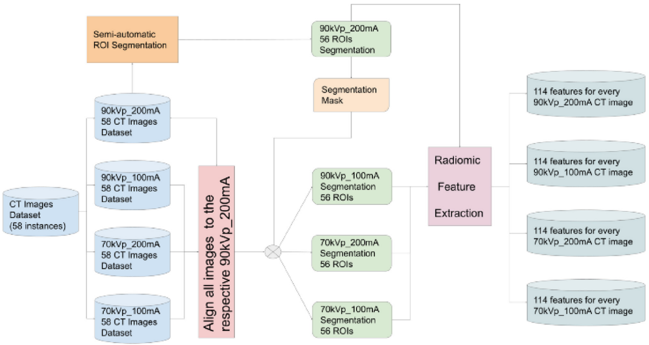

Fig. 1. Injury segmentation and Radiomic feature extraction process flowchart.

B. Injury Segmentation

A semi-automated segmentation approach of the ROIs was adopted and was initially used on all the 90 kVp_200 mA CT images. The segmentation of the injuries was performed on 3-D Slicer (version 4.10.2) using the WandEffect label tool [36]. The "Threshold Paint" option was selected and the thresholds for minimum and maximum gray level intensity values, after experimentation, were set at 100 and 1200, respectively. In case of an air bubble forming inside the skeletal muscle after the injury was induced, the air bubble was included in the ROI. The "maximum pixels per click" selection setting of the WandEffect label tool kit was set at 2000 pixels and the "Fill Volume" option was selected. Finally, manual correction of the segmented area had to be applied as well.

Consequently, all CT instances were aligned to the 90 kVp_200mA CT images and the masks created during the segmentation of the 90 kVp_200 mA CT images were then used as segmentation masks to extract the ROIs from the rest of the CT images of the different imaging parameters as well for all CT instances. Exceptions were two instances where the 90 kVp_200 mA CT images were rotated in relation to the CT images of different value combinations of imaging parameters. This resulted in slightly different ROIs to be extracted from the 90 kVp_200 mA CT images which would not allow the results of the different imaging parameters settings to be comparable and as a result these two instances were excluded from the dataset. In Fig. 1, the image processing flowchart is depicted, including the muscle injury segmentation stage.

C. Feature Extraction

Currently, there are two approaches to radiomic features extraction. The first approach uses mathematical models to extract features relating to imaging features, such as texture, intensity, or shape and is usually referred to as "feature-based" or "hand-crafted" radiomics [5]. The second is usually referred to as "deep learning-based radiomic (DLR) features" and is based on the hypothesis that once the ROI has been segmented accurately from a medical image by a deep neural network, the information about the segmented region is already stored within the network. Since the total number of available instances from the dataset would be considered limited for the purpose of DLR extraction, the "hand-crafted" feature extraction approach was selected. The open source radiomics platform PyRadiomics [37] was used for the extraction of hand-crafted radiomics features, that can be categorized into the following groups [5]. The follow-up day after the injury was included as well in the initial pool of features, resulting in a total of 114 features.

1) *Shape Features [38]:* Which provide quantitative description of geometric properties of the ROIs/VOIs, such as surface area, total volume, diameter, sphericity, or surface-to-volume ratio.
2) *First Order Statistics (Histogram-Based Features) [38]:* Which describe the fractional volume for the selected region of voxels and the distribution of the voxels' intensity, for example minimum, maximum, mean, variance, skewness, or kurtosis.
3) *Second Order Statistics (Textural Features) [5], [39]:* These features are extracted based on the following matrices derived from intensity relationships of neighboring voxels in a 3-D image.
    a) Gray Level Co-occurrence Matrix.
    b) Gray Level Run Length Matrix.
    c) Gray Level Size Zone Matrix.
    d) Neighboring Gray Tone Difference Matrix.
    e) Gray Level Dependence Matrix.

The shape "Voxel Volume" feature, which represents the volume of the skeletal muscle injury (ROI), was selected as the target value to predict for the ML models. Radiomics were extracted from all 58 instances forthfold, resulting in four separate datasets, one for each different value combination of the CT imaging parameters, as can be seen in Fig. 1.

D. Data Preprocessing

The range of the extracted radiomic feature values can vary greatly. Standardization is a technique often used as part of data preprocessing in an ML study when features of the input dataset have significant differences between their ranges. On distance-based ML algorithms, like support vector machines (SVMs) or k-nearest neighbors (k-NN), features with values that are of different ranges do not weigh the same when calculating distance. Standardization gives all features the same influence on the distance metric. Also, regressions like LASSO or Ridge that place a penalty on the magnitude of the coefficients associated to each variable can have deficient performance when fitting data with feature values of different variance.

Since in this study models based on support vector regressor (SVR), Ridge, and Lasso regressors are going to be created and tested, the StandardScaler implementation of the free software ML library Scikit-learn [40] was used to produce scaled data that has zero mean and unit variance.

On distance-based ML algorithms, like SVR, Ridge, and LASSO regressors, if the features of the input dataset have significant differences between their ranges, they do not have equal weight when calculating distance. Also, regressions like LASSO or Ridge that place a penalty on the magnitude of the coefficients associated to each variable can have deficient performance when fitting data with feature values of different variance. Since the range of the extracted radiomic values used in this study varies greatly, we used Standardization to harmonize influence on all features on the distance metric. Specifically, the StandardScaler implementation of the free software ML library Scikit-learn [40] was used to produce scaled data that has zero mean and unit variance.

E. Feature Selection

For the feature selection process, we tested a combination of feature selection techniques. Initially, we used two filter techniques, one supervised (mutual information) and one unsupervised (Pearson's correlation coefficient), in order to quickly minimize the dimensionality of the dataset and afterwards we proceed with three wrapper techniques (Backward Elimination, Forward Selection, and Bidirectional Elimination) to find the optimal feature combination for each ML algorithm, as described below. We also included three ML algorithms that use embedded feature selection methods in our set (LASSO, Ridge, and ElasticNet regressors).

Feature selection is a pivotal step in a radiomics studies' workflow, due to the high-dimensional dataset that radiomics feature extraction produces [42], [43], [44]. Although the number of extracted radiomics features can be exceptionally large, the usual case is that a lot of them can be either highly correlated to one another and/or irrelevant to the target value. Including such features can result in a model that is easy to overfit, noise sensitive and with reduced generalizability [41]. In short, by reducing the number of input features the data becomes more statistically significant.

As a first step in our feature selection process the concept of mutual information [47] was used to measure the mutual dependence between the target value and the rest of the extracted radiomic features. All features with a mutual information score of 0 toward the target value were excluded from the dataset as being independent of the target value and thus irrelevant to the task at hand. Consequently, Pearson's correlation coefficient [48] was utilized to measure the strength and direction of linear association between all pairs of remaining features. Highly correlated features with a Pearson's correlation coefficient value greater than 0.97 were compared with one another and only the features with the highest correlation to the target value were evaluated. The rest of the highly correlated features were considered redundant and were excluded from the dataset. As the last step in the feature selection process, three wrapper methods were used to search for the feature subset that leads to optimal predictive performance for each of the ML algorithms tested in the study. Wrapper

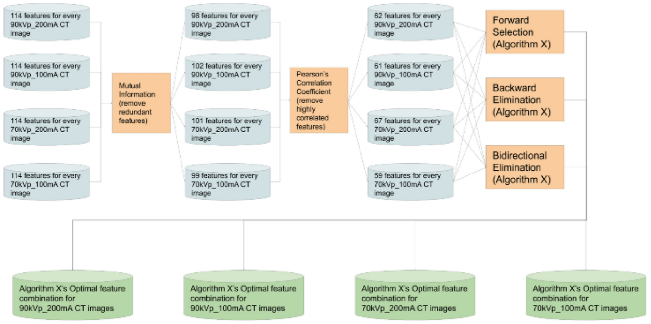
114 features for every 90kVp_200mA CT image
114 features for every 90kVp_100mA CT image
114 features for every 70kVp_200mA CT image
114 features for every 70kVp_100mA CT image
Mutual Information (remove redundant features)
98 features for every 90kVp_200mA CT image
102 features for every 90kVp_100mA CT image
101 features for every 70kVp_200mA CT image
99 features for every 70kVp_100mA CT image
Pearson's Correlation Coefficient (remove highly correlated features)
62 features for every 90kVp_200mA CT image
61 features for every 90kVp_100mA CT image
67 features for every 70kVp_200mA CT image
59 features for every 70kVp_100mA CT image
Forward Selection (Algorithm X)
Backward Elimination (Algorithm X)
Bidirectional Elimination (Algorithm X)
Algorithm X's Optimal feature combination for 90kVp_200mA CT images
Algorithm X's Optimal feature combination for 90kVp_100mA CT images
Algorithm X's Optimal feature combination for 70kVp_200mA CT images
Algorithm X's Optimal feature combination for 70kVp_100mA CT images

Fig. 2. Feature Selection process flowchart for Algorithm X.

methods [45], [46] search for an optimal feature subset fit to a specific ML algorithm by creating different subsets of features, building ML models based on the algorithm selected using the feature subsets and evaluating the models' performance on a chosen metric [49]. In this study, the metric that the models' performance was evaluated on was the mean absolute error (MAE), described in metrics section that follows.

Different strategies of creating the feature subsets result in different wrapper methods. For the purposes of this study, we used the following.

1) *Forward Selection:* The model starts empty, and features keep being added to it for as long as the models' performance keeps being improved. In each iteration the feature that gets added to the features subset is the one that leads to the greatest improvement of the models' performance.
2) *Backward Elimination:* The model starts with all the features and in each iteration, features keep being removed from it for as long as the models' performance keeps getting improved. In each iteration the feature that gets removed from the features subset is the one that leads to the greatest improvement of the models' performance.
3) *Bidirectional Elimination:* Works as Forward Selection, starting from an empty feature subset and adding the feature that leads to the greatest improvement of the models' performance in each iteration but with the possibility of deleting a feature that was previously added.

Since the values of the radiomics change depending on the different value combinations of the CT imaging parameters, the filter methods feature selection process had to be performed for each CT image dataset separately. Moreover, because wrapper methods are algorithm specific, the wrapper methods feature selection process had to be performed separately for each combination of CT imaging parameters and ML algorithm, as can be seen in Fig. 2. The optimal feature combination for any ML algorithm and any CT imaging parameters was selected according to the best performing model on the metric MAE.

F. Machine Learning Regression Algorithms

In total ten supervised ML regression algorithms were evaluated: 1) Least Squares Linear Regression [50]; 2) Ridge Regression [51]; 3) LASSO Regression [52]; 4) Elastic Net [53] Regression; 5) AdaBoost [54] regressor; 6) Gradient Boost [55] regressor; 7) eXtreme Gradient Boosting (XGBoost) [56] regressor; 8) Random Forest [57] regressor; 9) Decision Tree [58] regressor; and 10) SVR [59]. All the ML algorithm implementations used in this study can be found in the open-source software ML library Scikit-learn [40], except for XGBoost's implementation which can be found in the XGBoost open-source software library [60].

Linear Regression, being one of the most well-known and understood ML regression algorithms, can be used as a benchmark in this study. In addition to the ordinary least squares linear regression, three more implementations of the algorithm were tested which by the application of different regularization terms are enhancing the performance of linear regression in high-dimensional problems. LASSO and Ridge Regressions impose the L1 and L2 regularization to the cost function, respectively, while Elastic Net uses a linear combination of L1 and L2 regularization [61].

ML algorithms based on the concept of ensemble learning are considered the state-of the art solution when dealing with complex and high-dimensional data [62]. There are three main categories of ensemble learning algorithms: 1) bagging [63]; 2) boosting [64]; and 3) stacked generalization or stacking [65]. Embedded bagging and boosting ensemble learning regression algorithms were implemented in this study with Random Forest and AdaBoost, Gradient Boost and XGBoost, respectively.

G. Training and Evaluation

To train an ML model to predict the volume of the muscle injury of a given rat over time we need to rearrange our dataset into pairs of starting and ending instances. We will be referring to these pairs as snapshots. The ML model will be fed with the selected input features, including the initial "Voxel Volume," of the injury's starting instance as well as the time length in days between the initial and the target instance, and the output will be the "Voxel Volume" of the injury during the target instance.

Each of the three rats of the longitudinal study contributed with 15 snapshots of injuries. Two of the instances of rat no. 2 had to be excluded from the dataset because these 90 kVp_200 mA CT images were rotated in relation to CT images of different value combinations of imaging parameters and thus contributing with only six snapshots. As a result, the longitudinal study offered 36 snapshots, while each rat ($n = 20$) of the Single Post-injury study contributed with one snapshot, which led to a total of 56 snapshots of injuries per value combination of imaging parameters.

A variation of the leave one out cross validation (LOOCV) [66] method was used to evaluate the models' performance. LOOCV method was selected in order to counter the limited number ($n = 56$) of snapshots of starting and ending points of injuries per value combination of imaging parameters. The LOOCV method allows us to use more data on the training of our models than any other validation method. According to this method, our data are divided into two separate sets; 1) a training and 2) a validation set. The training set consists of all the snapshots, apart from the one snapshot which incorporates the validation set of each training iteration. So, only one snapshot is used for validation, and the rest of the dataset is used for the training of the model. This validation process will be repeated as many times as the total number of snapshots. This way we end up having a prediction of the Voxel Volume for each of the 56 snapshots. In our variation of the LOOCV method, when the snapshot of the validation set is part of the longitudinal study, the rest of the snapshots that were produced by CT images of the same rat were excluded from the training set as well, to prevent the introduction of bias.

H. Metrics

To compare the predictive performance of the ML models we computed the following performance measures.

1) *MAE* is the average of the absolute errors of the model's predictions against the snapshots' target values.
2) *RMSE* is the square root of the average of the squared errors of the model's predictions against the snapshots' target values.
3) *R-squared (R2)* or coefficient of determination represents the proportion of the variance of the target value explained by the input features in a regression model.

MAE and RMSE are scale dependent, so they can be used to compare the performance of different predictive regression models for a particular dataset but not between datasets [67]. Smaller MAE and/or RMSE values indicate better predictive performance, while larger R2 values indicate better fit of the data with values ranging from 0 to 1.

Since, according to literature [68], MAE is the more natural measure of average error magnitude, and that, unlike RMSE, it is unambiguous, it was used as the primary model performance measure in this study for performance comparison and optimization purposes.

I. Hyperparameter Optimization

Hyperparameter optimization or tuning is the process of finding a set of hyperparameter values which allow an ML algorithm to better fit the data achieving the best possible performance according to a predefined metric, MAE in this case, on a cross validation set. Hyperparameter optimization plays a vital role in the prediction accuracy of ML algorithms. Different automatic hyperparameter optimization search algorithms have been proposed, such as grid search [69], random search [69], Bayesian search [70], gradient-based search [71], and multifidelity search [72] methods.

In this study, we applied the implementation of Bayesian optimization available in the open-source software ML library Scikit-Optimize [73] on the best performing algorithms after the feature selection process was completed. Table A2 (supplementary material) lists the sets and the ranges of the hyperparameters per ML algorithm that were optimized with the use of Bayesian optimization. Bayesian optimization was selected due to its ability to achieve comparable improvement of the predictive performance of ML algorithms in significantly reduced runtime compared with other optimization methods [74].

J. Ensemble Learning

As a final step in our methodology, we implemented ensemble learning techniques on the results of the ML regression algorithms to further improve our predictions. Ensemble learning [75] refers to the process of developing a single "strong" ML model that solves a computational problem by strategically combining multiple differently performing "weaker" ML models, treating them as a "committee" of solvers. The principle is that the prediction of the committee, when individual predictions are combined appropriately, should have better overall accuracy than any individual model (committee member). After the completion of the hyperparameter optimization process, we used the outputs of the four best performing models (XGBoost, Ridge regression, Gradient Boost, and Random Forest) of the 70 kVp, 100 mA dataset to create weighted average ensemble learning models.

Weighted average or weighted sum ensemble [76] is an ensemble learning approach that combines predictions from multiple models, where the contribution of each model is weighted proportionally to the model's predictive ability. That weight is then multiplied by the model's prediction and is used for the calculation of the average prediction. In regression, the average prediction is calculated using the arithmetic mean, as shown in following equation:

$$P_e = \frac{\sum_{i=1}^{n} w_i \times P_i}{\sum_{i=1}^{n} w_i}$$

where P_e is the prediction of the ensemble, n is the total number of predictors contributing to the ensemble, P_i is the prediction of predictor i, w_i is the weight assigned to predictor i.

We tested three different ensemble combinations, starting with the best performing model and progressively adding models according to best performance. We also used six different approaches for the assignment of weight to the ML models where each model contributes.

1) Equally ($w = 1$) to the prediction.
2) By $w = 1/\text{MAE}$ to the prediction.
3) By $w = 1/\text{RMSE}$ to the prediction.
4) By $w = R^2$ to the prediction.
5) According to the model's MAE performance to the prediction. Weights get values from 1 to the number of models.
6) According to the model's RMSE performance to the prediction. Weights get values from 1 to the number of models.

The sixth approach of weight assignment gave the best results. In Table II, we present the performance metrics of all the ensemble combinations for the sixth weight assignment approach.

III. Results

In this section, we evaluate the trained models' ability to predict the volume of skeletal muscle injury in rats over time for the two sets of imaging parameters. In Table I, we can see the performance metrics of the six best performing algorithms after the optimization process for the 70 kVp, 100 mA

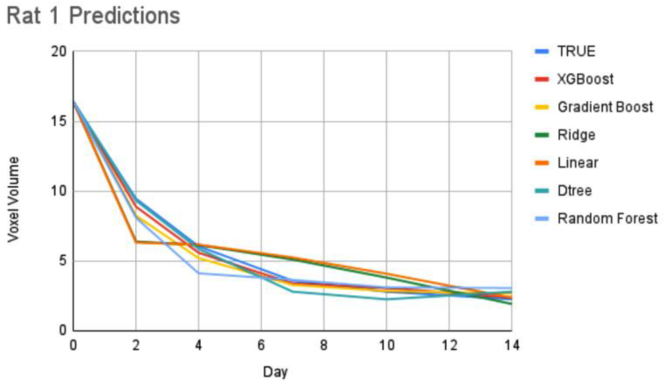

Fig. 3. Diagram of the Voxel Volume predictions of the six best models over a period of 14 days for the Longitudinal's study Rat 1.

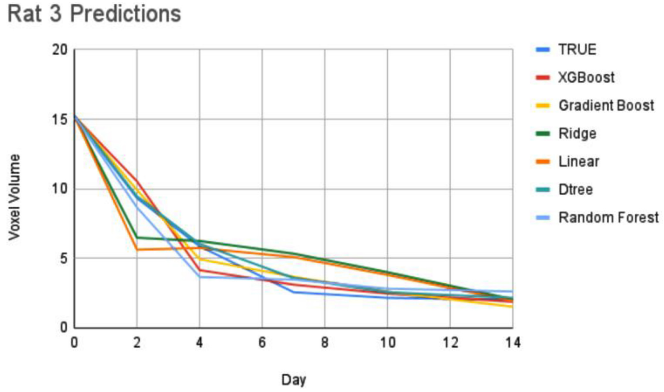

Fig. 4. Diagram of the Voxel Volume predictions of the six best models over a period of 14 days for the Longitudinal's study Rat 3.

TABLE I
BEST PERFORMING MODELS AFTER THE OPTIMIZATION PROCESS PER IMAGING PARAMETERS COMBINATION

Imaging Parameters	MAE	RMSE	R^2
70kV, 100mA	XGBoost 1.34	Ridge 2.42	Ridge 0.80
90kV, 100mA	XGBoost 1.51	Linear Reg 2.50	Linear Reg 0.79
70kV, 200mA	XGBoost 1.74	XGBoost 2.91	XGBoost 0.72
90kV, 200mA	XGBoost 1.57	XGBoost 2.63	XGBoost 0.77

dataset. The model that achieves the best MAE score of 1.336 uses the XGBoost algorithm on the 70 kVp, 100 mA dataset and uses seven input features, six of which are radiomics features plus the follow-up day from initial injury. It is notable that the best R^2 score is achieved with Ridge regression for the same imaging parameter dataset, but with double number of features (15) although its main principle leads on penalizing and minimizing the feature space. Figs. 3 and 4 depict the Voxel Volume predictions of the six best models over a period of 14 days for the longitudinal's study Rats 1 and 2, respectively.

Individual diagrams for the best performing models per ML algorithm can be seen in Figs. A1–A12 (supplementary material), while the input features of the six best performing models for the 70 kVp, 100 mA dataset can be seen in Supplementary Material in Table A3 in the supplementary material. The performance metrics of the best performing models after the conclusion of the feature selection process of the ten ML algorithms that were tested in this study for each of the different imaging parameters datasets can be seen in Tables A4–A7 of the supplementary material. The best performing models highlighted in Tables A4–A7, in the supplementary material, with green, were subjected to Bayesian hyperparameter optimization. The metrics of the best models after the optimization process are depicted in Tables A8–A11 of the supplementary material.

TABLE II
METRICS OF THE BEST PERFORMING WEIGHTED AVERAGE ENSEMBLE MODELS FOR THE 70 kV, 100 mA IMAGING PARAMETERS COMBINATION

Ensembles	MAE	RMSE	R^2
XGB	1.336	2.469	0.796
XGB+Ridge	1.298	**2.174**	0.842
XGB+Ridge+GB	**1.220**	2.178	**0.846**
XGB+Ridge+GB+RF	1.240	2.244	0.840

Table II shows the performance of the applied ensemble learning using weighted average among the four best models for the same imaging parameters case of 70 kVp and 100 mA. One can clearly see that performance gets enhanced when XGBoost is combined with Ridge regression and more learners (Gradient Boost and Random Forest) with all metrics exhibiting significant improvement compared to the best performing individual ML model (XGBoost). Finally, the best performing model presents to be the XGB+Ridge+GB ensemble model, with the lowest mean average error and the highest R^2 value of all. RMSE had a very similar value for both XGB+Ridge ensemble cases, with and without GB (2.178 and 2.174, respectively).

IV. DISCUSSION

The results revealed that the predictions of the injury's volume of the XGBoost and Gradient Boost models follow very closely the recovery trend of the muscle injury in the longitudinal study's Rats 1 and 3, as seen in Figs. 3 and 4 (also in Figs. A1–A4 of the supplementary material), contrary to the rest algorithms and for example Ridge regression (Figs. A5 and A6 in the supplementary material), that appears to achieve comparable performance to XGBoost (the best performing model).

Hyperparameter optimization improved model performance up to ~20% depending on the case and the metric. R^2 presented the smaller improvement in most cases.

As seen in Table I, even though models were trained on datasets of different CT imaging parameters combinations, all best performing models were trained on datasets where the tube current was set to 100 mA, indicating an advantage in comparison to 200 mA. Moreover, four out of the six best performing models were trained on the dataset where the peak kilovoltage was set to 70 kVp indicating an advantage in comparison to 90 kVp. The best individual predictive model was based on the XGBoost Regressor algorithm and was trained using seven input features, six of which were radiomics extracted from the 70 kVp, 100 mA dataset. This model achieved an MAE score of 1.336, an RMSE score of 2.469, and an R^2 score of 0.796. Ensemble learning for the same imaging parameters led to improved model performance when combining XGBoost with Ridge regression and slightly with GB too, but exhibited worse performance when Random Forest was also included in the ensemble procedure, as seen in Table II. This verifies that ensemble learning reaches a plateau in the improving performance depending on the involved algorithms and their inner-working variation.

The best achieved RMSE value of 2.174 (XGB+Ridge ensemble model) that corresponds to 6.7% of the mean initial injury volume, shows indeed a big improvement (3.1 times better performance) compared to the exponential model of Paun et al. on the same dataset who achieved RMSE of 6.8 in their calculations.

The small size of the cohort ($n = 23$) which led to a limited ($n = 56$) number of snapshots is the main limitation of the current study. However, a comparison was applied on the different ML algorithms for the best performing model for skeletal muscle injury healing process based on the evaluation of hand-crafted radiomic features. Due to the small cohort size, the proposed methodology should be further validated with larger datasets.

V. Conclusion

In this study, we developed a methodology to apply ML techniques on radiomics extracted from contrast enhanced CT images of skeletal muscle injuries in rats to develop ML models to predict the injury recovery progress over time in rats. Our results suggest that radiomics can successfully be used to predict the volume of a skeletal muscle injury in rats over time. Moreover, our results show that different CT imaging parameter settings for tube current and peak kilovoltage impact the predictive performance of the ML regression models, indicating that lower values of tube current and peak kilovoltage contribute to more accurate predictions.

As further steps, multi-institutional studies on larger cohorts and different animal species should be conducted to further validate and standardize our methodology. Applications of our methodology and/or findings could be used as a tool to assist clinicians on skeletal muscle injury diagnosis and treatment, through the prediction of the unassisted recovery progress. Following the study of Contreras-Muñoz et al. [35] on the development of a surgical model of skeletal muscle injury in rats that reproduces human sports lesion, our work can also have a direct impact on human studies. Our aim is to further investigate ML models for human translation of the predictions Finally, the complete methodology proposed in this study can be implemented in different applications (beyond oncology and skeletal muscle injuries) using the relevant CT imaging data.

Acknowledgment

This publication reflects only the author's views, and the European Union is not liable for any use that may be made of the information contained therein. The authors declare that they have no known conflicts of interest in terms of competing financial interests or personal relationships that could have an influence or are relevant to the work reported in this article.

References

[1] E. Bercovich and M. C. Javitt, "Medical Imaging: From roentgen to the digital revolution, and beyond," *Rambam Maimonides Med. J.*, vol. 9, no. 4, Oct. 2018, Art. no. e0034, doi: 10.5041/RMMJ.10355.

[2] R. J. Gillies, P. E. Kinahan, and H. Hricak, "Radiomics: Images are more than pictures, they are data," *Radiology*, vol. 278, no. 2, pp. 563–577, Feb. 2016, doi: 10.1148/radiol.2015151169.

[3] P. Lambin et al., "Radiomics: Extracting more information from medical images using advanced feature analysis," *Eur. J. Cancer*, vol. 48, no. 4, pp. 441–446, Mar. 2012, doi: 10.1016/j.ejca.2011.11.036.

[4] J. E. van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler "Radiomics in medical imaging—'How-to' guide and critical reflection," *Insights Imag.*, vol. 11, p. 91, Dec. 2020. [Online]. Available: https://doi.org/10.1186/s13244-020-00887-2

[5] P. Papadimitroulas et al., "Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization," *Phys. Med.*, vol. 83, pp. 108–121, Mar. 2021. [Online]. Available: https://doi.org/10.1016/j.ejmp.2021.03.009

[6] Z. Li, Y. Wang, J. Yu, Y. Guo, and W. Cao, "Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma," *Sci. Rep.*, vol. 7, no. 1, p. 5467, Jul. 2017, doi: 10.1038/s41598-017-05848-2.

[7] X. Yao and Y. Liu, "Machine learning," in *Search Methodologies*, E. K. Burke and G. Kendall, Eds. Boston, MA, USA: Springer, 2005. [Online]. Available: https://doi.org/10.1007/0-387-28356-0_12

[8] B. Koçak, E.Ş. Durmaz, E Ateş, and Ö. Kılıçkesmez, "Radiomics with artificial intelligence: A practical guide for beginners," *Diagn. Interv. Radiol.*, vol. 25, no. 6, pp. 485–495, Nov. 2019, doi: 10.5152/dir.2019.19321.

[9] S. Bernatz et al., "Comparison of machine learning algorithms to predict clinically significant prostate cancer of the peripheral zone with multiparametric MRI using clinical assessment categories and radiomic features," *Eur. Radiol.*, vol. 30, no. 12, pp. 6757–6769, Dec. 2020. [Online]. Available: https://doi.org/10.1007/s00330-020-07064-5

[10] S. Yao, H. Jiang, and B. Song, "Radiomics in prostate cancer: Basic concepts and current state-of-the-art," *Chin. J. Acad. Radiol.*, vol. 2, no. 3, pp. 47–55, Feb. 2020. [Online]. Available: https://doi.org/10.1007/s42058-019-00020-3

[11] F. Bianconi, I. Palumbo, A. Spanu, S. Nuvoli, M. L. Fravolini, and B. Palumbo, "PET/CT radiomics in lung cancer: An overview," *Appl. Sci.*, vol. 10, no. 5, p. 1718, Mar. 2020. [Online]. Available: https://doi.org/10.3390/app10051718

[12] I. Fornacon-Wood, C. Faivre-Finn, J. P. B. O'Connor, and G. J. Price, "Radiomics as a personalized medicine tool in lung cancer: Separating the hope from the hype," *Lung Cancer*, vol. 146, pp. 197–208, Jun. 2020. [Online]. Available: https://doi.org/10.1016/j.lungcan.2020.05.028

[13] Y. Zhou et al., "The application of artificial intelligence and radiomics in lung cancer," *Precis. Clin. Med.*, vol. 3, no. 3, pp. 214–227, Sep. 2020. [Online]. Available: https://doi.org/10.1093/pcmedi/pbaa028

[14] X. Yi et al., "Computed tomography radiomics for predicting pathological grade of renal cell carcinoma," *Front. Oncol.*, vol. 10, Jan. 2021, Art. no. 570396, doi: 10.3389/fonc.2020.570396.

[15] M. Kocher, M. I. Ruge, N. Galldiks, and P. Lohmann, "Applications of radiomics and machine learning for radiotherapy of malignant brain tumors," *Strahlentherapie Onkologie*, vol. 196, pp. 856–867, Oct. 2020. [Online]. Available: https://doi.org/10.1007/s00066-020-01626-8

[16] S. Bae et al., "Robust performance of deep learning for distinguishing glioblastoma from single brain metastasis using radiomic features: Model development and validation," *Sci. Rep.*, vol. 10, Jul. 2020, Art. no. 12110. [Online]. Available: https://doi.org/10.1038/s41598-020-68980-6

[17] Y. S. Sung, B. Park, H. I. Park, and S. S. Lee, “Radiomics and deep learning in liver diseases,” *J. Gastroenterol. Hepatol.*, vol. 36, no. 3, pp. 561–568, Mar. 2021. [Online]. Available: https://doi.org/10.1111/jgh.15414

[18] H. Wenmo, Y. Huayu, X. Haifeng, and M. Yilei, “Radiomics based on artificial intelligence in liver diseases: Where are we?” *Gastroenterol. Rep.*, vol. 8, no. 2, pp. 90–97, Apr. 2020. [Online]. Available: https://doi.org/10.1093/gastro/goaa011

[19] S. Arnaldo et al., “Handcrafted MRI radiomics and machine learning: Classification of indeterminate solid adrenal lesions,” *Magn. Reson. Imag.*, vol. 79, pp. 52–58, Jun. 2021. [Online]. Available: https://doi.org/10.1016/j.mri.2021.03.009

[20] F. Torresan et al., “Radiomics: A new tool to differentiate adrenocortical adenoma from carcinoma,” *BJS Open*, vol. 5, no. 1, Jan. 2021, Art. no. zraa061. [Online]. Available: https://doi.org/10.1093/bjsopen/zraa061

[21] E. Guerriero, L. Ugga, and R. Cuocolo, “Artificial intelligence and pituitary adenomas: A review,” *Artif. Intell. Med. Imag.*, vol. 1, no. 2, pp. 70–77, Aug. 2020, doi: 10.35711/aimi.v1.i2.70.

[22] Y. Zhang et al., “Radiomics approach for prediction of recurrence in nonfunctioning pituitary macroadenomas,” *Front. Oncol.*, vol. 10, Dec. 2020, Art. no. 590083, doi: 10.3389/fonc.2020.590083.

[23] I. Cetin et al., “Radiomics signatures of cardiovascular risk factors in cardiac MRI: Results from the U.K. biobank,” *Front. Cardiovasc. Med.*, vol. 7, Nov. 2020, Art. no. 591368, doi: 10.3389/fcvm.2020.591368.

[24] Q. Feng and Z. Ding, “MRI radiomics classification and prediction in Alzheimer’s disease and mild cognitive impairment: A review,” *Curr. Alzheimer Res.*, vol. 17, no. 3, pp. 297–309, Mar. 2020. [Online]. Available: https://doi.org/10.2174/1567205017666200303105016

[25] Y. W. Park et al. “Radiomics features of hippocampal regions in magnetic resonance imaging can differentiate medial temporal lobe epilepsy patients from healthy controls,” *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, Art. no. 19567. [Online]. Available: https://doi.org/10.1038/s41598-020-76283-z

[26] A. F. Leite, K. D. F. Vasconcelos, H. Willems, and R. Jacobs, “Radiomics and machine learning in oral healthcare,” *Prot. Clin. Appl.*, vol. 14, no. 3, May 2020, Art. no. 1900040. [Online]. Available: https://doi.org/10.1002/prca.201900040

[27] E. K. Oikonomou, M. Siddique, and C. Antoniades, “Artificial intelligence in medical imaging: A radiomic guide to precision phenotyping of cardiovascular disease,” *Cardiovasc. Res.*, vol. 116, no. 13, pp. 2040–2054, Nov. 2020.

[28] Q. Chen, T. Xia, M. Zhang, N. Xia, J. Liu, and Y. Yang, “Radiomics in stroke neuroimaging: Techniques, applications, and challenges,” *Aging Dis.*, vol. 12, no. 1, pp. 143–154, Feb. 2021, doi: 10.14336/AD.2020.0421.

[29] V. Romeo and S. Maurea, “The new era of advanced placental tissue characterization using MRI texture analysis: Clinical implications,” *EBioMedicine*, vol. 51, Jan. 2020, Art. no. 102588, doi: 10.1016/j.ebiom.2019.11.049.

[30] B. Paun et al., “Modelling the skeletal muscle injury recovery using in vivo contrast-enhanced micro-CT: A proof-of-concept study in a rat model,” *Eur. Radiol. Exp.*, vol. 4, p. 33, Jun. 2020, doi: 10.1186/s41747-020-00163-4.

[31] S. Wong, A. Ning, C. Lee, and B. T. Feeley, “Return to sport after muscle injury,” *Curr. Rev. Musculoskelet. Med.*, vol. 8, no. 2, pp. 168–175, Jun. 2015, doi: 10.1007/s12178-015-9262-2.

[32] N. J. Gibbs, T. M. Cross, M. Cameron, and M. T. Houang, “The accuracy of MRI in predicting recovery and recurrence of acute grade one hamstring muscle strains within the same season in Australian rules football players,” *J. Sci. Med. Sport*, vol. 7, no. 2, pp. 248–258, 2004. [Online]. Available: https://doi.org/10.1016/s1440-2440(04)80016-1

[33] M. Kumaravel, P. Bawa, and N. Murai, “Magnetic resonance imaging of muscle injury in elite American football players: Predictors for return to play and performance,” *Eur. J. Radiol.*, vol. 108, pp. 108–164, Nov. 2018, doi: 10.1016/j.ejrad.2018.09.028.

[34] J. Ekstrand, J. C. Healy, M. Waldén, J. C. Lee, B. English, and M. Hägglund, “Hamstring muscle injuries in professional football: The correlation of MRI findings with return to play,” *Brit. J. Sports Med.*, vol. 46, no. 2, pp. 112–117, Feb. 2012, doi: 10.1136/bjsports-2011-090155.

[35] P. Contreras-Muñoz et al., “A new surgical model of skeletal muscle injuries in rats reproduces human sports lesions,” *Int. J. Sports Med.*, vol. 37, no. 3, pp. 183–190, Mar. 2016, doi: 10.1055/s-0035-1555933.

[36] “Documentation/nightly/modules/editor.” Accessed: Oct. 12, 2022. [Online]. Available: https://www.slicer.org/wiki/Documentation/Nightly/Modules/Editor

[37] J. J. M. van Griethuysen et al., “Computational radiomics system to decode the radiographic phenotype,” *Cancer Res.*, vol. 77, pp. 104–107, Nov. 2017. [Online]. Available: https://doi.org/10.1158/0008-5472.can-17-0339

[38] V. Kumar et al., “Radiomics: The process and the challenges,” *Magn. Reson. Imag.*, vol. 30, no. 9, pp. 1234–1248, Nov. 2012, doi: 10.1016/j.mri.2012.06.010.

[39] V. Parekh and M. A. Jacobs, “Radiomics: A new application from established techniques,” *Expert Rev. Precis. Med. Drug Develop.*, vol. 1, no. 2, pp. 207–226, Mar. 2016, doi: 10.1080/23808993.2016.1164013.

[40] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011, doi: 10.48550/arxiv.1201.0490.

[41] P. Afshar, A. Mohammadi, K. N. Plataniotis, A. Oikonomou, and H. Benali, “From handcrafted to deep-learning-based cancer radiomics: Challenges and opportunities,” *IEEE Signal Process. Mag.*, vol. 36, no. 4, pp. 132–160, Jul. 2019, doi: 10.1109/MSP.2019.2900993.

[42] Z. C. Lipton, “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,” *Queue*, vol. 16, no. 3, pp. 31–57, May 2018, doi: 10.1145/3236386.3241340.

[43] M. J. Kearns, *Computational Complexity of Machine Learning*. Cambridge, MA, USA: MIT Press, 1990.

[44] L. Chen, “Curse of dimensionality,” in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. Boston, MA, USA: Springer, 2009. [Online]. Available: https://doi.org/10.1007/978-0-387-39940-9_133

[45] R. Kohavi and G. John, “Wrappers for feature subset selection,” *Artif. Intell.*, vol. 97, nos. 1–2, pp. 273–324, Dec. 1997. [Online]. Available: https://doi.org/10.1016/S0004-3702(97)00043-X

[46] T. N. Lal, O. Chapelle, J. Weston, and A. Elisseeff, “Embedded methods,” in *Feature Extraction, Studies in Fuzziness and Soft Computing*, vol. 207, I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Berlin, Germany: Springer, pp. 137–165, doi: 10.1007/978-3-540-35488-8_6.

[47] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 4, pp. 623–656, Oct. 1948, doi: 10.1002/j.1538-7305.1948.tb00917.x.

[48] W. Kirch, “Pearson’s correlation coefficient,” in *Encyclopedia of Public Health*. Dordrecht, The Netherlands: Springer, 2008, pp. 1090–1091. [Online]. Available: https://doi.org/10.1007/978-1-4020-5614-7_2569

[49] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, “Data transformations,” in *Data Mining*, 4th ed., I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Eds. Burlington, MA, USA: Morgan Kaufmann, 2017, pp. 285–334. [Online]. Available: https://doi.org/10.1016/B978-0-12-804291-5.00008-8

[50] T. L. Lai, H. Robbins, and C. Z. Wei, “Strong consistency of least squares estimates in multiple regression II,” *J. Multivariate Anal.*, vol. 9, pp. 343–361, Sep. 1979.

[51] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased estimation for nonorthogonal problems,” *Technometrics*, vol. 12, no. 1, pp. 55–67, 2012, doi: 10.1080/00401706.1970.10488634.

[52] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Royal Stat. Soc. B, Stat. Methodol.*, vol. 58, no. 1, pp. 267–288, 1996. [Online]. Available: http://www.jstor.org/stable/2346178

[53] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” *J. Royal Stat. Soc. B, Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[54] R. E. Schapire, “Explaining AdaBoost,” in *Empirical Inference*, B. Schölkopf, Z. Luo and V. Vovk, Eds. Berlin, Germany: Springer, 2013. [Online]. Available: https://doi.org/10.1007/978-3-642-41136-6_5

[55] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *Ann. Stat.*, vol. 29, pp. 1189–1232, Oct. 2001, doi: 10.1214/AOS/1013203451.

[56] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Disc. Data Min.*, New York, NY, USA, 2016, pp. 785–794. [Online]. Available: https://doi.org/10.1145/2939672.2939785

[57] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[58] J. R. Quinlan, “Induction of decision trees,” *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986. [Online]. Available: https://doi.org/10.1007/BF00116251

[59] M. Awad and R. Khanna, “Support vector regression,” in *Efficient Learning Machines*. Berkeley, CA, USA: Apress, Apr. 2015, pp. 67–80. [Online]. Available: https://doi.org/10.1007/978-1-4302-5990-9_4

[60] “XGBoost documentation.” Accessed: Oct. 13, 2022. [Online]. Available: https://xgboost.readthedocs.io/en/latest/index.html#

[61] A. Torang, P. Gupta, and D. J. Klinke, “An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets,” *BMC Bioinformat.*, vol. 20, no. 1, p. 433, Aug. 2019. [Online]. Available: https://doi.org/10.1186/s12859-019-2994-z

[62] F. Huang, G. Xie, and R. Xiao, “Research on ensemble learning,” in *Proc. Int. Conf. Artif. Int. Comput. Intell.*, Nov. 2009, pp. 249–252, doi: 10.1109/AICI.2009.235.

[63] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996. [Online]. Available: https://doi.org/10.1023/A:1018054314350

[64] L. Breiman, “Arcing classifier (with discussion and a rejoinder by the author),” *Ann. Stat.*, vol. 26, no. 3, pp. 801–849, 1998, doi: 10.1214/aos/1024691079.

[65] D. H. Wolpert, “Stacked generalization,” *Neural Netw.*, vol. 5, no. 2, pp. 241–260, 1992. [Online]. Available: https://doi.org/10.1016/S0893-6080(05)80023-1

[66] C. Sammut and G. I. Webb, “Leave-one-out cross-validation,” in *Encyclopedia of Machine Learning*. Boston, MA, USA: Springer, 2011, pp. 600–601. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_469

[67] R. J. Hyndman and A. B. Koehler, “Another look at measures of forecast accuracy,” *Int. J. Forecast.*, vol. 22, no. 4, pp. 679–688, 2006, doi: 10.1016/j.ijforecast.2006.03.001.

[68] C. Willmott and K. Matsuura, “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance,” *Clim. Res.*, vol. 30, no. 1, pp. 79–82, Dec. 2005, doi: 10.3354/CR030079.

[69] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 281–305, Mar. 2012. [Online]. Available: https://jmlr.csail.mit.edu/papers/volume13/bergstra12a/bergstra12a.pdf

[70] A. Agnihotri and N. Batra. “Exploring Bayesian optimization.” Distill. May 2020. [Online]. Available: https://distill.pub/2020/bayesian-optimization/

[71] C. B. Do, C. S. Foo, and A. Y. Ng, “Efficient multiple hyperparameter learning for log-linear models,” in *Proc. NIPS*, Dec. 2007, pp. 377–384.

[72] Y.-Q. Hu, Y. Yu, W.-W. Tu, Q. Yang, Y. Chen, and W. Dai, “Multi-fidelity automatic hyper-parameter tuning via transfer series expansion,” *AAAI*, vol. 33, no. 1, pp. 3846–3853, Jul. 2019. [Online]. Available: https://doi.org/10.1609/aaai.v33i01.33013846

[73] “Scikit-optimize.” Accessed: Oct. 14, 2022. [Online]. Available: https://scikit-optimize.github.io/stable/

[74] J. Wu, X.-Y. Chen, H. Zhang, L.-D. Xiong, H. Lei, and S.-H. Deng, “Hyperparameter optimization for machine learning models based on Bayesian optimization,” *J. Electron. Sci. Technol.*, vol. 17, pp. 26–40, Mar. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1674862X19300047

[75] J. Kittler and F. Roli, “Multiple classifier systems,” in *Proc. 1st Int. Workshop MCS*, Cagliari, Italy, 2000, pp. 981–986.

[76] M. Shahhosseini, G. Hu, and H. Pham, “Optimizing ensemble weights and hyperparameters of machine learning models for regression problems,” 2019, *arXiv:1908.05287*.