

Feature Selection in Healthcare Datasets: Towards a Generalizable Solution<sup>☆</sup>Ida Maruotto<sup>a</sup>, Federica Kiyomi Ciliberti<sup>a</sup>, Paolo Gargiulo<sup>a,b,1</sup>, Marco Recenti<sup>a,b,\*,1</sup><sup>a</sup> Institute of Biomedical and Neural Engineering, Reykjavik University, Reykjavik, Iceland<sup>b</sup> Department of Science, Landspítali University Hospital, Reykjavik, Iceland

## ARTICLE INFO

Dataset link: <https://biavrsea.is>, <https://restore-project.ru.is>

## Keywords:

Feature selection  
Artificial intelligence  
Machine learning  
Dimensionality reduction  
Healthcare dataset  
Scalability  
Biomedical signals

## ABSTRACT

**Background and objective:** The increasing dimensionality of healthcare datasets presents major challenges for clinical data analysis and interpretation. This study introduces a scalable ensemble feature selection (FS) strategy optimized for multi-biometric healthcare datasets aiming to: address the need for dimensionality reduction, identify the most significant features, improve machine learning models' performance, and enhance interpretability in a clinical context.**Methods:** The novel waterfall selection, that integrates sequentially (a) tree-based feature ranking and (b) greedy backward feature elimination, produces as output several sets of features. These subsets are then combined using a specific merging strategy to produce a single set of clinically relevant features. The overall method is applied to two healthcare datasets: the biosignal-based BioVRSea dataset, containing electromyography, electroencephalography, and center-of-pressure data for postural control and motion sickness assessment, and the image-based SinPain dataset, which includes MRI and CT-scan data to study knee osteoarthritis.**Results:** Our ensemble FS approach demonstrated effective dimensionality reduction, achieving over a 50% decrease in certain feature subsets. The new reduced feature set maintained or improved the model classification metrics when tested with Support Vector Machine and Random Forest models.**Conclusion:** The proposed ensemble FS method retains selected features essential for distinguishing clinical outcomes, leading to models that are both computationally efficient and clinically interpretable. Furthermore, the adaptability of this method across two heterogeneous healthcare datasets and the scalability of the algorithm indicates its potential as a generalizable tool in healthcare studies. This approach can advance clinical decision support systems, making high-dimensional healthcare datasets more accessible and clinically interpretable.

## 1. Introduction

The rapid growth of available data in modern society, boosted by recent advancements in artificial intelligence (AI) and data mining, presents nowadays a major challenge in managing high-dimensional datasets, leading to the need for dimensionality reduction and feature selection (FS) techniques. The issue of high dimensionality, also known as the “curse of dimensionality”, presents significant challenges in data analysis, model training, and interpretation, as datasets with numerous features are more prone to sparsity and noise, leading to possible overfitting affecting the performances [1,2].

FS techniques enable the identification of relevant and informative features, reducing the dimensionality of datasets while preserving essential information and enhancing computational efficiency. This process not only aids in model interpretability, a crucial factor in domains

such as finance and healthcare, but also helps to optimize resource use, particularly when computational or data collection resources are limited [3].

FS methods are mainly divided into filter and wrapper categories [4]. In the first case, features are selected based on performances on various statistical tests. Person's correlation, ANOVA, and chi-square are among the most used and efficient filter techniques [5]. The wrapper methods, computationally more expensive, choose the relevant features by performing tests on a subset of features in a loop, training them with a selected algorithm adding or removing variables from the tested subset. These include forward selection, backward elimination, and Recursive Feature Elimination (RFE) [6]. When the quality of filter and wrapper methods are merged, the so-called embedded methods are generated [7]. Furthermore, the technique of principal component

<sup>☆</sup> This research was funded by RANNIS, Icelandic Research Fund (Grant of Excellence no: 239612-051) and by European Union (Horizon 2024 - CORDIS Grant agreement ID: 101057778).

\* Correspondence to: Institute of Biomedical and Neural Engineering, Reykjavik University, Menntavegur 2, 102 Reykjavik, Iceland.

E-mail address: [marcor@ru.is](mailto:marcor@ru.is) (M. Recenti).

<sup>1</sup> These authors share last authorship.

analysis (PCA) is often performed to reduce the number of features, creating a new set of artificial features, reducing the number, and, at the same time, preserving the maximum amount of information [8].

In healthcare modern studies a vast amount of biomedical data can be generated by various sources, including biosensors and medical imaging, creating a demand for efficient analysis tools to support proper AI and machine learning (ML) models' applications. The proper set of healthcare human-derived data is crucial for training efficiently predictive models presenting a complex task to optimize the clinical valuable outcomes while minimizing time and cost [9,10]. The identification and the retaining of the most meaningful features can enhance the accuracy and robustness of the predictive models, which is crucial in the healthcare domain where decisions can have a direct impact on patient outcomes and on the reaching of precise clinical and diagnostic evaluations [11].

When working with healthcare dataset that includes detailed bio-signal analysis, such as ECG, EEG, and EMG data, FS techniques are proven essential. These signals might often generate irrelevant or redundant features that obscure the most valuable information and allow for more precise clinical interpretability of the data.

For instance, in the classification of cardiac conditions using ECG data, FS algorithms such as RFE and genetic algorithms have been demonstrated as effective in improving ML model performances [12, 13]. Wrapper and embedded methods were also performed efficiently to improve automatic biometric human recognition using ECG [14].

Similarly, in EEG-based studies for neurological disorders, methods like sequential backward selection [15] or simple decision tree-based selection [16] have been employed to reduce the number of features while maintaining or improving classification accuracy. Li et al. [17] experimented multiple FS techniques to detect mild depression in young university students with 128 channels EEG data finding the optimal performances with techniques based on correlation and greedy forward/backward selection.

When machine learning is involved in the feature selection process, architectures can be very complex. Xue et al. [18] proposed SWD-NAS, which leverages a dual-attention module and architectural weight normalization to improve fairness and performance during the search process in differentiable architecture search. Mechanism that can also be used in generative adversarial networks to improve convergence [19].

Recently, biology-inspired FS techniques like the Artificial Bee Colony algorithm [20] and the Binary Particle Swarm Optimization (PSO) [21,22] have been applied effectively to eliminate superfluous time and frequency domain features for EMG signal classification outperforming the classification models that used the whole set of extracted features. PSO has also been used on Hepatitis B virus diagnosis, obtaining a small classification rate [23]. Also, multi-objective feature selection has been widely used in evolutionary computation to perform feature selection [24,25].

In the context of healthcare, combining filter and wrapper algorithms for FS in embedded hybrid approaches resulted particularly performing. For instance, Alzubi et al. [26] proposed a fusion of a filter and a wrapper method outperforming the basic FS methodologies to select Single Nucleotide Polymorphisms for rare diseases classification, while Pavithra and Jayalakshmi [27] focused on cardiovascular diseases classification combining Random Forest (RF), ADA-Boost and Pearson Coefficient FS methods. Xue et al. [28] proposed a importance-guided particle swarm optimization based on MLP (IGPSO) for feature selection purposes, on public datasets including medical. Also, on 11 tumor datasets, EAR-FS can effectively select 3% of the features by keeping the classification accuracy over 93% [29]. Medical imaging, another healthcare domain with large and clinical informative data generation, also relies heavily on optimized FS [30]. For instance, deep learning-based techniques and PCA were performed to reduce dimensionality while preserving essential diagnostic information: these methods have shown clear improvement in the field of

image-based oncology allowing for example, more precise breast and prostate diagnosis [31,32].

Within the knee cartilage degeneration studies using MRI and CT scan medical images, the SinPain dataset has been demonstrated to be efficient in distinguish healthy and degenerative cartilages using radiomics data together with densitometric and geometric features related to bones, cartilage and muscles [33]. In specific, FS techniques that considered both statistical data analysis and correlation analysis were performed on the full cohort of 98 subjects. Two approaches were used: firstly, two highly correlated features were merged creating a single new feature, alternatively, one of the two correlated features was removed, basing the decision on the least significant feature according to statistical analysis [34]. Moreover, a ML study was conducted on radiomic data extracted from cartilages (on both image modalities CT and MRI) [35]. FS was here performed through correlation analysis, where features with a correlation greater than 90% were removed. The FS process, guided by both correlation and feature importance analyses, identified the texture and shape-related features as key predictors in distinguishing between healthy and degenerative knee cartilage classes.

FS techniques have been recently applied to predict adaptive responses during postural control tasks with the BioVRSea paradigm. The study employs multiple physiological signals, including Electromyography (EMG), Electroencephalography (EEG), and Center of Pressure (CoP) crucial for understanding and diagnosing balance disorders and neurodegenerative diseases, such as Parkinson's disease, and for evaluating susceptibility to motion sickness [36,37]. Various feature selection (FS) methods, including filter and wrapper techniques, were applied to BioVRSea feature sets for classification purposes. Specifically, Center of Pressure (CoP) and EMG features were used to classify healthy and unhealthy lifestyles. FS filters, such as ANOVA and mutual information tests, highlighted CoP as the most effective signal for distinguishing between these groups [38]. Additionally, applying ANOVA to EEG, EMG, and CoP data achieved high accuracy in predicting phases of postural adaptation, underscoring the complementary roles of muscle activity and pressure in balance control [39]. Furthermore, EEG, CoP, and basic heart rate features, reduced via PCA, could distinguish individuals with a history of concussion from healthy controls, though interpretability remained limited due to PCA's artificial output [40]. Heart rate features, selected through a backward elimination wrapper FS, effectively predicted motion sickness, with movement-phase data indicating a stronger symptom response and suggesting possible cardiovascular irregularities in healthy subjects during complex postural tasks [41]. These studies have mainly focused on single bio-signals and never considered the complete set of available features, amounting to a dataset approximately five times the number of subjects. Each of these features holds potential for clinical interpretability, both individually and within specific sub-groups.

### 1.1. Goal of the present study

In healthcare applications, a significant amount of highly meaningful features are continuously generated from human sources, whether they are derived from biosignals or medical imaging, creating a large availability of potential clinical important features. It is crucial to understand which of these features retain the most impactful clinical meaning through reliable feature selection methods. This understanding is essential not only for machine learning applications, which rely on accurate and efficient models, but more importantly for clinicians who need to interpret these variables in their decision-making.

In this study, we propose an ensemble feature selection strategy grounded in sequential tree-based feature ranking (Fig. 1). We apply this strategy to two datasets with features extracted from both biosignals and medical images, using data from the BioVRSea and Sinpain datasets. This study introduces a feature selection workflow that aims to:

- identify the most significant features in medical datasets;

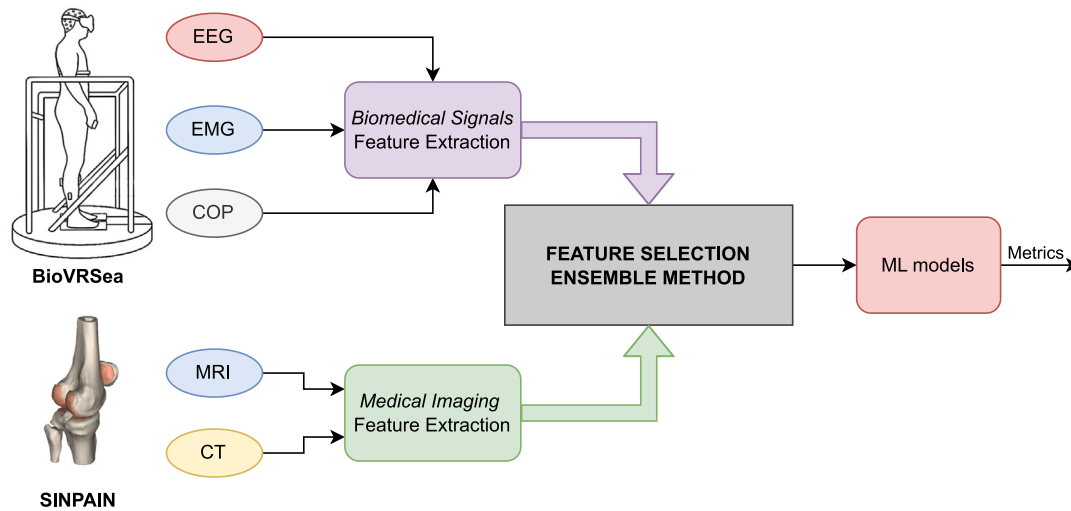


Fig. 1. Workflow of the present research.

- improving the performance of predicting models;
- enhance interpretability of results in clinical contexts;
- ensure computational efficiency of feature selection.

## 2. Materials and methods

The novel ensemble features selection approach is applied to two different datasets described below: the BioVRSea, a postural control dataset based on biosignals features, and the SinPain database, which includes imaging features from knee MRI and Ct-Scan.

The BioVRSea studies involving human participants were reviewed and approved by the Icelandic National Bioethics Committee (Application no: 17-183-S1). All the patients/participants provided their written informed consent to participate in the study. The Sinpain study was approved by the Icelandic Bioethics Commission (approval number: VSN-19-050V1). Prior to any scans, subjects provided comprehensive written consent, ensuring transparency and accountability. To ensure patient privacy, all data are anonymized and securely stored on a dedicated hospital server approved by the Data Protection Unit.

### 2.1. BioVRSea dataset

The BioVRSea project [42], firstly introduced in 2020, at the Institute of Biomedical and Neural Engineering (IBNE) at Reykjavik University, is an advanced research paradigm designed to study human postural control and balance behavior using virtual reality (VR) combined with a dynamic platform that simulates sea waves [43]. This novel approach aims to enhance the understanding of the neurophysiological processes behind balance and postural adaptation, particularly in a complex sensory environment that challenge stability, such as being on a simulated moving boat. The study employs multiple physiological signals, including Electromyography (EMG), Electroencephalography (EEG), and Center of Pressure (CoP). Therefore, the BioVRSea setup is a so-called multi-metric system that combines a moving platform and virtual reality to assess postural control with multiple biosignals, during a 6 phases experiment.

In the first phase (baseline), the subject views a static mountain environment while the platform remains still, and EEG and EMG recordings begin [36,37]. The following phases simulate the virtual experience of being on a small boat at sea starting with no platform movement (PRE phase - only visual waves input), then increases movement in three intervals of 40 s each (25%, 50%, and 75% of maximal machine

amplitude), then finish with again only visual input and no movement of the platform re-doing the PRE phase (POST phase). The subjects are instructed to balance themselves in the PRE and POST phases without holding on to the bar. Demographic data and the Motion Sickness Questionnaire (MSSQ) were collected before the experiment, and afterward to assess differences between the pre- and post-experimental phases. The focus was on the main motion sickness symptoms like nausea, dizziness, sweating, and salivation identified by Golding [44].

Approximately 500 subjects, both healthy and with diseases that affect normal postural control behavior, have been measured so far with the BioVRSea paradigm. However, the dataset of healthy subjects was resized to 388 based on the questionnaire and signal availability. The classification target label for the ML models in this work, “BioVRSea Effect Index”, is binary and refers to all the self-assessed motion sickness symptoms analyzed. A subject is labeled as “sick” (1) if any symptoms increased after the experiment compared to before; otherwise, the label is “not sick” (0). Among the many categorical features included in the questionnaire, two are here selected as input features: gender and Lifestyle Index. The latter is calculated as binary index considering Body Mass Index, nicotine, caffeine and alcohol consume and physical activity frequency.

The biosignals are not analyzed raw but several features are extracted from them [37]. The EEG signals were recorded using a 64-channel cap (sampling at 4096 Hz) and after preprocessing (including artifact removal and signal interpolation), power spectral density (PSD) was computed for each electrode and merged into 5 regions (parietal, occipital, frontal, temporal left, and right). Relative PSD values were calculated for each region in different brain bands (alpha, beta, gamma, delta, and theta), this resulted in 150 EEG features (5 regions  $\times$  6 phases  $\times$  5 bands). EMG data were recorded from 6 calf muscles, three on each leg. After a preprocessing with a Butterworth filter, spectral parameters were calculated using the discrete Fourier transform. Time-domain and frequency-domain features were then extracted resulting in a total of 43. Considering all the phases, the total of the EMG features is 1548. CoP measurements were taken using sensors under each foot platform. A Savitsky-Golay filter was applied to smooth the signal during the preprocessing, and features were extracted based on medio-lateral and anteroposterior sways. A total of 175 features were calculated (35 features  $\times$  5 phases).

In the dataset, some biosignals are missing for some subjects. This led to slightly different dataset sizes depending on the considered features: 322 samples in EEG data, 324 in COP data, 336 in EMG data, and only 259 complete subjects. In Table 1 is reported the training set composition with respect to the inspected feature subset: it is relative to 80% of the full dataset, but keeping the same ratio between classes of the original dataset.

**Table 1**  
BioVRSea training set composition with respect to the inspected feature subset.

| Algorithm execution       | BioVRSea sick | BioVRSea not sick |
|---------------------------|---------------|-------------------|
| EEG                       | 209 (65%)     | 113 (35%)         |
| EMG                       | 221 (66%)     | 115 (34%)         |
| COP                       | 210 (65%)     | 114 (35%)         |
| ALL                       | 167 (64%)     | 92 (36%)          |
| ALL + Gender (+Lifestyle) | 166 (64%)     | 91 (36%)          |

## 2.2. SinPain dataset

The RESTORE (<http://restoreproject.eu/>) and SINPAIN (<https://www.osteoarthritis-sinpain.eu/>) projects are two EU-funded projects focused on studying knee cartilage degeneration. In particular, RESTORE goal was to develop advanced nanobiomaterials and bioprinting technologies for treating knee chondral lesions, while SINPAIN aims to develop new strategies for knee osteoarthritis (OA), which is among the most common joint disorders worldwide. As part of these projects, IBNE recruited subjects diagnosed with OA (degenerative class, ‘D’) and healthy volunteers (control class, ‘C’) to acquire a specific set of 275 images. RESTORE/SINPAIN dataset includes 136 computed tomography (CT) and 139 Magnetic Resonance Imaging (MRI) scans acquired at Landspítali University Hospital in Reykjavik, following standardized image acquisition and patient positioning protocols. Several features were extracted from the processing of these images, including densitometric and geometric features related to bones, cartilage and muscles, and radiomic features.

The SINPAIN dataset contains the CT and MRI of 98 subjects, for a total of 144 scanned legs, where 56 subjects scanned both left and right legs. From the MRI, femoral, lateral tibial, medial tibial, and patellar cartilages were manually segmented and then reported onto the CT through the process of image registration. All the steps in segmentation, registration, and feature extraction were performed using the medical software MIMICS (Materialise Interactive Medical Image Control System, Materialise, Belgium). This validated workflow, described in details in [33,45], allows to extract radiomic, densitometric, and geometric features from different anatomical tissues.

The radiomic features were extracted from both the CT-scan registered cartilages and the MRI segmented ones, using the PyRadiomics library [46]. They include first-order statistics, shape-based features, texture ones (based on gray-value levels), and features derived from the wavelet transform of the images, creating two datasets, one for CT and one for MRI. The total radiomic features are 3245 from the CT images and 3368 from the MRI images [47]. As proposed in [35], a correlation analysis was performed to refine the dataset, and features with a correlation greater than 80% were removed, reducing the feature set to 420 for the CT and 581 for the MRI. Moreover, from CTs, densitometric and geometric muscular and fat tissue features are extracted by selecting a small cylinder containing them from the axial view with the virtual histology technique [48,49]. The virtual histology was applied on eleven muscles, including the vastus medialis, intermedius and lateralis, the biceps femoris, the gastrocnemius lateral and the medial head. Then, muscular and adipose tissue were differentiated based on density thresholds in Hounsfield units (HU). The average density, standard deviation (SD), and volume of each tissue were computed, for a total of 66 features (3 features for each of the 2 extracted tissues of the eleven muscles). Finally, from the four registered cartilages, volumes, surfaces, and radiodensity (average and SD) were computed for a total of 16 cartilage features [34].

The SINPAIN data extracted for the current study come from 102 legs with 4 classes of features:

- Cartilages: it contains the densitometric and geometric features extracted by the CT registered cartilages.
- Muscles: it contains the densitometric and geometric features from the muscle cylinder segmented on the CT.

- Radiomics MRI: it contains the radiomic features extracted by the cartilages segmented on the MRI.
- Radiomics CT: it contains the radiomic features extracted by the cartilages registered on the CT.

As ML target label, people with a diagnosis of OA are considered sick while all the others are considered to belong to the healthy class.

## 2.3. Feature selection ensemble method

The proposed FS ensemble method approach presented in this study, which full architecture is visible in [Fig. 2] including input and output, is split into two main sections: a first waterfall selection block and a next feature merging block from which the final set of features is obtained. The final set of features is then the input for the classification models from which the evaluation metrics are computed. The inputs of the pipeline are the dataset itself and 50 unique seeds ( $\in S$ , where  $S$  is the seed set); as a matter of fact, the minimum number of seeds to have a good generalization is 20 as demonstrated by Colas et al. [50].

To prevent bias in seed selection, each seed was generated using a random number generator initialized with a unique seed, simulating the randomness of unfixed execution. As stated before, the number of seeds has been selected to guarantee a good generalization and avoid effects from seed distribution.

The dataset is split into two subsets namely the 80% training set and the 20% test set. The employment of a fixed seed value for the split guarantees the repeatability and the reliability of the entire experimentation process. Moreover to maintain identical class ratios in training and test set and to avoid an arbitrary division of the dataset we applied a stratification splitting technique.

Before providing the data to the pipeline, we performed a scaling procedure using the Z-score normalization [51], where  $x$  is the feature,  $\mu$  is the training set mean value and  $\sigma$  is the training set SD (see Eq. (1)). Z-score normalization is a well-known technique used to transform the features to have the properties of a standard normal distribution with a mean of 0 and an SD of 1. Z-score normalization is typically associated with normally distributed data, but it can be calculated for non-normal distributions in the same way. However, the interpretation can be more challenging because the percentiles derived from the scores may not correspond to standard normal probabilities. Besides, the use of normalization in the scope of this work is to boost the convergence and performance of AI models, and not give probabilistic interpretations of the normalized features, so this is a neglectable drawback.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

The scaling algorithm’s parameters (means and SDs) have been calculated based on the training set, and then used to scale the test set without re-calculation, in order to allow a test for generalizability.

### 2.3.1. Section 1: Waterfall selection

The novel waterfall features selection section can be considered the core of the methodology proposed.

Its structure, inspired by software deployment waterfall model [52], is composed of two parts, composing the waterfall, and they are called *feature selection A* and *feature selection B*. This approach aims to perform both exploration (with FS A) and exploitation (with FS B) of the searching space [Fig. 3], and retrieving both outputs at the end of the process. The selected ML algorithm is the Gradient Boosting (GB) [53]; it builds weak classifiers sequentially, preferably trees. Each weak classifier learns from the residual errors of its predecessors, and each tree’s predictions are weighted based on their accuracy. Usually, tree-based models are the most explainable and preferable for clinical applications. Moreover, it has been considered that Recursive feature elimination based on GB between the most effective one in healthcare data [54].

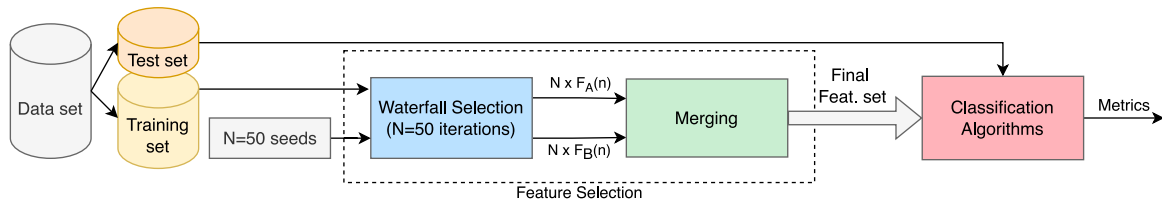


Fig. 2. System's General Architecture.

FS A consists of a basic selection of the most significant features from the GB model and the output set is called  $F_A$ . This first method aims to select a subset of features from the dataset, focusing on the most contributing ones in the prediction process. The selection is performed on the trained model, and the importance of each feature concerning the target variable is evaluated on both the performance and the prediction error. Using a threshold for the parameter selection is common practice because by introducing a regularization factor, the computational resources can focus on the most relevant features. This model simplification leads to faster training, improved performance, and reduced risk of overfitting [55]. Indeed, in this work, an L1 (LASSO) penalty [56] is used with the corresponding threshold of  $10^{-5}$ . All the features below this threshold are discarded, obtaining a virtual filter, and to do so, it adds a penalty term proportional to the absolute values of the coefficient of features.

The FS B is a *Backward Sequential FS*, on a reduced version of the dataset comprehensive of only the significant features included in the FS A. The set of features obtained from this process is called  $F_B$ .

The backward FS is a greedy algorithm that works stepwise and operates systematically to identify the most significant features. It has been proven to be effective with datasets with an unbalanced sample number versus features [57]. The sequential FS process starts with constructing a predictive model using all the available features in the dataset, which will be used as a baseline for successive iterations. In this case, the model's effectiveness is evaluated based on its accuracy. Still, any other metric can be chosen to evaluate how well the model predicts the target variable.

The backward selection is done so that reducing the number of features should not decrease by more than 1% at a time. This is because variations of such a small scale are not very significant with respect to the relevance of the extrapolation of fewer features. However, the Backwards FS is a time-consuming algorithm, so to avoid using too many computational resources to perform a backward passage, the maximum number of features obtainable from the first FS is fixed to 50. It is not a dynamic threshold depending on the number of features, still, we have found that it can be a good trade-off between accuracy, computational time, clinician relevance, and a few features that a human can consider at a glance.

The FS A is performed on the entire input training set, while the FS B is performed with 5-fold cross-validation (CV), because FS A is intended as a rough exploration and FS B is an exploitation of the space [58,59]. Considering each seed  $n \in S$ , where  $S$  is the seed set, the FS A output  $F_A(n)$  is a list of features and their weights in the model. Instead, FS B receives  $F_A(n)$  as input and retrieves an unordered list of features as output,  $F_B(n)$ . Considering all the sets,  $\forall n \in S$ , we can derive the supersets  $F_A$  and  $F_B$ , containing every algorithm execution.

In Pseudocode 1 a high level description of the novel waterfall selection is provided.

### 2.3.2. Section 2: Feature merging

The pipeline is executed 50 times, with each random generator set to the same unique fixed seed for every iteration. Thus each iteration can produce a different set of results. This led to the necessity of merging them as homologous sets [Fig. 4].

### Pseudocode 1 Waterfall selection

```

function WATERFALL( $n\_seeds$ , dataset)
   $i \leftarrow n\_seeds$ 
  while  $i \neq 0$  do
    TRAIN(GB, dataset)
    importances  $\leftarrow$  COMPUTE_FEATURE_IMPORTANCES(GB)
     $F_{A_i} \leftarrow$  SORT(importances BY importance DESCENDING, LIMIT
max_feat_number)
     $F_{B_i} \leftarrow$  BACKWARDFEATURESELECTION(GB, dataset,  $F_{A_i}$ )
     $n\_seeds \leftarrow n\_seeds - 1$ 
  end while
   $F_A \leftarrow$  UNION_ALL( $F_{A_i}$  for each  $i$ )
   $F_B \leftarrow$  UNION_ALL( $F_{B_i}$  for each  $i$ )
end function

```

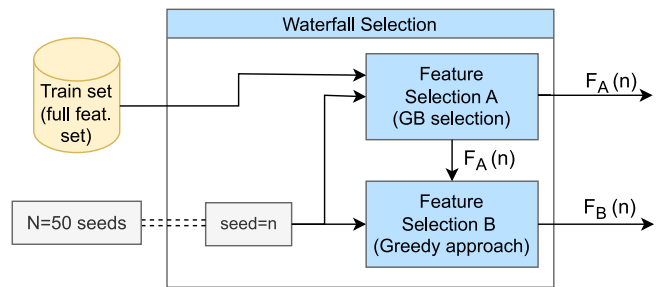


Fig. 3. Waterfall Selection section of the algorithm.

The final  $F_R$  set is obtained by ranking the features  $F_A$  according to the number of executions they have been considered important and taking the 75 quartiles of the set. We can call this the *repetition* method because it ranks the features based on how many times they are repeated among the 50 executions.

The  $F_P$  set is formed by collecting all the features that were found to be useful at any different iteration by FS B,  $F_B$ . This is referred to as the *performance* method because it is based on the assumption of not losing performances during the greedy reduction.

The  $F_W$  set is determined by evaluating the features  $F_A$  according to the sum of their weights during the 50 executions. These features are then ranked according to this sum, and next, the top 75% are included in the set.

When merging, the final set  $F$  appears as the set of the common features to  $F_R$ ,  $F_P$ , and  $F_W$ , namely repetition, performance, and weighted sets.

We selected the threshold 75% to enable further FS while ensuring that essential features are not eliminated. After some iterations, we selected it as a trade-off between feature amount and overfitting. The workflow of how the merging works is described in Pseudocode 2.

### 2.3.3. Classification algorithms

The algorithm was validated on two distinct datasets (BioVRSea and SINPAIN), addressing different classification targets(self-reported score and disease assessment). In BioVRSea tests, the pipeline has been

**Pseudocode 2** Feature merging

```

FRsubset ← SORT(  $F_A$  BY repetition_num DESCENDING)
 $F_R$  ← FILTER( FRsubset ≥ PERCENTILE(FRsubset.importance, 75))
FWsubset ← SORT(  $F_A$  BY weights DESCENDING)
 $F_W$  ← FILTER( FWsubset ≥ PERCENTILE(FWsubset.weights, 75))
 $F_P$  ←  $F_B$ 
for all feature IN features do
  if feature ∈  $F_R, F_W, F_P$  then
    APPEND feature TO final_set
  end if
end for

```

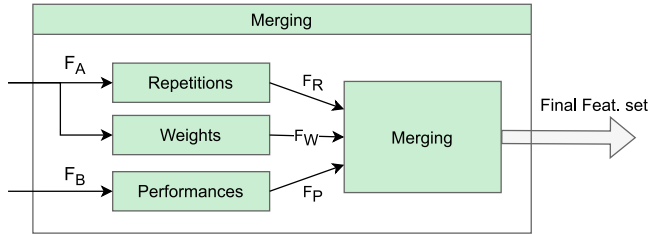


Fig. 4. Merging section of the algorithm.

executed separately on each biosignal analyzed (EMG, COP, and EEG) and their ensemble in the BioVRSea effect classification. In SINPAIN tests, the pipeline has been executed separately for each feature category (cartilages, muscles, radiomics CT and radiomics MRI) and their ensemble for OA classification. In both cases, the goal is to understand if the pipeline is more effective on homogeneous features or if it can be data-agnostic, keeping the same performances.

Moreover, both an SVM model and a Random Forest (RF) model are trained on the original feature sets and the reduced ones  $F$  to compare performances [Fig. 5]. The choice of these algorithms tries to span among different “explainable” architectures to test the effectiveness of FS.

RF algorithm [60] belongs to the ensemble learning category. It searches for the best-selecting features among a randomly selected subgroup of features and it builds multiple decision trees. Each tree gives its prediction to obtain the predicted class or value, and a majority vote selects the final class.

The Support Vector Machine (SVM) [61] tries to find the best hyperplane separating two classes by maximizing the distance between the hyperplane and the nearest data points. To handle non-linear decisions, SVM uses a method known as the kernel trick, which transforms the input features into a higher dimensional space using functions like polynomial or Radial Basis Function (RBF) kernels. Since SVM is sensitive to the choice of the kernel function, we selected the RBF, shaped like a Gaussian, thanks to its property of dataset dimension invariance [62].

Sometimes, the use of categorical features when building the model can worsen the models’ performances, the so-called absent level [63]; thus, in both cases, the pipeline has been tested with and without categorical features in order to evaluate FS capabilities. The performances have been evaluated in CV on 5-folds.

### 2.3.4. Evaluation metrics

The metric considered for the classification evaluation is the F1 score. It represents the harmonic mean of precision and recall [64], providing a balanced measure between the two. Precision (P) is defined as the ratio of true positives (TP) to the sum of true positives and false positives (FP), while recall (R), also known as sensitivity, is the ratio of true positives to the sum of true positives and false negatives (FN). F1 score is particularly important in our case due to class imbalance

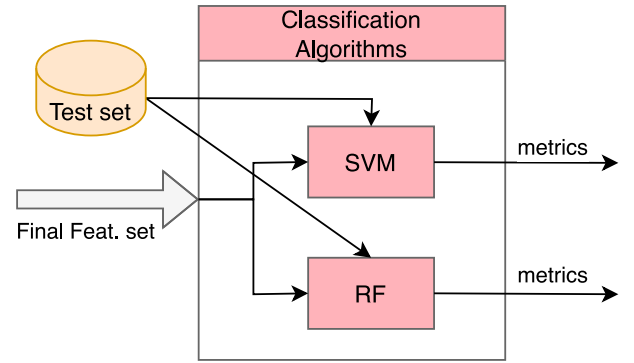


Fig. 5. Classification models.

**Table 2**

Feature number reduction in BioVRSea and SINPAIN datasets.

| Algorithm execution                      | Starting features | Selected features |
|------------------------------------------|-------------------|-------------------|
| BioVRSea effect EEG                      | 150               | 86 (57%)          |
| BioVRSea effect EMG                      | 1548              | 357 (23%)         |
| BioVRSea effect COP                      | 175               | 97 (55%)          |
| BioVRSea effect ALL                      | 1873              | 433 (23%)         |
| BioVRSea effect ALL + Gender             | 1873              | 437 (23%)         |
| BioVRSea effect ALL + Gender + Lifestyle | 1875              | 439 (23%)         |
| SINPAIN cartilages                       | 16                | 8 (50%)           |
| SINPAIN muscles                          | 66                | 32 (48%)          |
| SINPAIN radiomics MRI                    | 581               | 237 (41%)         |
| SINPAIN radiomics CT                     | 420               | 150 (36%)         |
| SINPAIN complete                         | 1082              | 341 (32%)         |

in both datasets. Metrics such as accuracy can be misleading in imbalanced settings, as they may be overly influenced by the majority class. The use of precision and recall allows the analysis of performances on each classes, and F1 score has the ability to combine the two in only one metric easier to compare. The F1 score can be calculated using Eq. (2), which ranges from 0 to 1, where a score of 1 indicates perfect precision and recall. This metric is particularly useful when class distributions are imbalanced and accuracy alone may be misleading.

$$F1 = 2 \times \frac{P \times R}{P + R} = 2 \times \frac{TP}{2 \times TP + FP + FN} \quad (2)$$

Thus the evaluated metrics are accuracy (ratio of correctly classified samples), precision, recall, and F1 score as defined by Hicks et al. [64] for medical AI applications. While we also computed other metrics (e.g., precision, recall, and accuracy), we prioritized the F1 score in this work for its ability to reflect the trade-off between false positives and false negatives, which is critical in clinical applications. All other evaluation metrics can be found in the Supplementary Materials.

## 3. Results

SVM and RF models are trained in CV on the original datasets and the reduced datasets obtained from the proposed FS process. The BioVRSea dataset consists of 1875 features, primarily comprising EMG signals, EEG data, and COP measurements, which are represented in similar quantities. In contrast, the dataset includes only two categorical features, making them the least represented (Table 2). The *gender* and *lifestyle* categorical features allowed us to carry out two different tests, with only one and with both, and check how well the algorithm can deal with them. In the SINPAIN dataset, the features are 1082: the most present feature group is the radiomics, mainly from MRI images, completely overshadowing the other categories composed of only 66 and 16 features.

The performances reported in Tables 3 and 5 are not directly derived from the FS approach. “Before feature selection” refers to the model’s performance when trained on the original dataset, which

**Table 3**

CV (k = 5) f1-score, before and after feature selection with the BioVRSea dataset.

| Feature set              | RF          |             | SVM         |             |
|--------------------------|-------------|-------------|-------------|-------------|
|                          | Full set    | Final set   | Full set    | Final set   |
| EEG                      | 0.75 ± 0.03 | 0.78 ± 0.01 | 0.74 ± 0.03 | 0.79 ± 0.01 |
| EMG                      | 0.76 ± 0.02 | 0.78 ± 0.03 | 0.78 ± 0.03 | 0.79 ± 0.01 |
| COP                      | 0.74 ± 0.03 | 0.78 ± 0.01 | 0.73 ± 0.03 | 0.78 ± 0.01 |
| ALL                      | 0.76 ± 0.03 | 0.77 ± 0.02 | 0.78 ± 0.03 | 0.79 ± 0.01 |
| ALL + Gender             | 0.75 ± 0.04 | 0.79 ± 0.02 | 0.78 ± 0.02 | 0.78 ± 0.01 |
| ALL + Gender + Lifestyle | 0.75 ± 0.04 | 0.78 ± 0.02 | 0.78 ± 0.02 | 0.79 ± 0.01 |

**Table 4**

Test f1-score after feature selection with BioVRSea dataset.

| Feature set              | RF   | SVM  |
|--------------------------|------|------|
| EEG                      | 0.79 | 0.78 |
| EMG                      | 0.80 | 0.79 |
| COP                      | 0.70 | 0.78 |
| ALL                      | 0.76 | 0.77 |
| ALL + Gender             | 0.80 | 0.79 |
| ALL + Gender + Lifestyle | 0.80 | 0.79 |

includes all the features. “After feature selection” refers to the model’s performance when trained on the reduced dataset, which contains only the features chosen by the FS process. The performances of the two models are F1 scores on a 5-fold CV, and we consider them as mean and SD. These results can be analyzed alongside the test set performances on unseen data. Finally, in the Supplementary material, Table S13 reports the minimum, maximum and average execution time per seed, as it highly depends on the convergence speed and number of iterations from FS B.

### 3.1. BioVRSea

As stated before, the FS was performed independently on EMG, EEG, and COP features. Moreover, the FS has been performed on a comprehensive dataset of all biosignals together with single categorical features. Table 1 shows how much the different executions may differ in dataset size. Besides, the different signals are actually quite different from each other, indeed, from a ML point of view, they have in common only the fact of being recorded simultaneously and the same target variable.

The EEG training set comprised 113 subjects who did not experience the influence of the experiment and 209 who did. EEG features are originally 150, and the FS extracted 86, corresponding to 57%. In this case, the number of features is lower than the dataset size, and the performances did not change when the feature set was reduced, so in this case fewer features do not lead to loss of information.

The EMG training set comprised 115 subjects not influenced by the experiment, and 221 were influenced. This reduction is the most effective, from 1548 features to only 357, corresponding to 23% of the original dataset.

The COP training set was made of 114 subjects who did not suffer, whereas 210 suffered from the experiment. Starting from 175 features, the algorithm extracted only 97 of them, the 55%.

The developed algorithm is not restricted to homogeneous data and can be applied to heterogeneous data as a unified dataset. Thus, the performances are evaluated on the complete dataset of every biosignal, obtaining a reduction from 1873 to 433 (23%), where the dataset is composed of 167 people who had symptoms and 92 who had not. The cross validation results are reported in Table 3, while the test set results are shown in Table 4. The feature set *F* is composed of 73.9% of features from EMG set, 14.3% from CoP set and 11.8% from EEG and the phases are equally represented.

When adding a categorical feature, the Gender, it changes the composition of the final dataset even though in the selected features it is not included. The EMG features rise to 74.4%, the EEG and COP are equally

**Table 5**

CV (k = 5) f1-score, before and after feature selection with the SINPAIN dataset.

| Feature set   | RF          |             | SVM         |             |
|---------------|-------------|-------------|-------------|-------------|
|               | Full set    | Final set   | Full set    | Final set   |
| Cartilages    | 0.67 ± 0.13 | 0.73 ± 0.08 | 0.66 ± 0.12 | 0.80 ± 0.07 |
| Muscles       | 0.62 ± 0.15 | 0.65 ± 0.12 | 0.58 ± 0.10 | 0.67 ± 0.11 |
| Radiomics MRI | 0.75 ± 0.19 | 0.83 ± 0.11 | 0.85 ± 0.08 | 0.92 ± 0.06 |
| Radiomics CT  | 0.75 ± 0.16 | 0.81 ± 0.10 | 0.75 ± 0.17 | 0.86 ± 0.07 |
| Complete      | 0.79 ± 0.14 | 0.82 ± 0.11 | 0.86 ± 0.12 | 0.94 ± 0.07 |

**Table 6**

Test f1-score after feature selection with the SINPAIN dataset.

| Feature set   | RF   | SVM  |
|---------------|------|------|
| Cartilages    | 0.83 | 0.81 |
| Muscles       | 0.79 | 0.70 |
| Radiomics MRI | 0.95 | 0.95 |
| Radiomics CT  | 0.88 | 0.90 |
| Complete      | 0.82 | 0.94 |

present with 12.8%. When adding the second categorical feature, the features’ group importance equal the execution without the categorical feature. Although the number of features selected and their overall composition do not change a lot, inspecting the selected features, they highly differ from the selection without categorical features.

Supplementary material, Tables S1, S2, and S3 report CV accuracy, precision, and recall on the BioVRSea dataset. Tables S7, S8, and S9 show test set accuracy, precision, and recall on the BioVRSea dataset.

### 3.2. SINPAIN

The pipeline has been fed by a total of 102 samples, 41 belonging to the positive class and 61 to the negative one. The positive class refers to people diagnosed with OA. One positive sample is missing in the radiomics CT dataset, leading to the training of 60 samples from the healthy class and 41 degenerate in the complete dataset. However, the ratio between classes is kept, and this difference can be neglected.

As done before, the performances are evaluated in CV (see Table 5) and on the test set (see Table 6).

Considering the complete dataset, the final composition of the dataset is 52.2% from radiological MRI features, 38.4% CT features, and 7.9% from muscle features, 1.5% cartilage features. Corresponding to 178 MRI features, 131 CT, 27 muscle, and 5 cartilage ones.

Tables S4, S5, and S6 in the Supplementary material report CV accuracy, precision, and recall on the SINPAIN dataset. Tables S10, S11, and S12 show test set accuracy, precision, and recall on the SINPAIN dataset.

## 4. Discussion

The objective of the work is to reduce the feature number for clinical evaluations. BioVRSea dataset is a collection of features extracted from biosignals, SINPAIN dataset is a collection of features extracted from images. Considering these characteristics, they are deeply different, but they share the fact that they are dealing with health-related data and fit into the scope of this work.

FS is considered effective when it identifies the most informative features that provide the most relevant data for the task. In ML terms, this means selecting the smallest possible feature subset that still allows the model to achieve accurate classification. In other words, the goal is to retain only the essential features that contribute significantly to the model’s performance while discarding any that do not add value or could even introduce noise. AI-driven approaches can also focus on important characteristics that physicians would otherwise neglect because of the high amount of information that can be misleading for diagnosis. On the one hand, in postural control analysis, FS is

important in order to finally reach clinical integration. On the other, in the SINPAIN dataset, the amount of information that can be extracted from an image sometimes complicates creating a metric system for clinical assessment of cartilage.

After FS, AI algorithms' performance may worsen. Still, it is neglectable with respect to the reduction amount, which is at least 43% in our tests, as shown in Table 2. However, this possibility never arises in this work because, as Table 3 and Table 5 show, the performances never worsen after FS; they increase or remain unchanged. Even with the least effective network, the OA classification using muscles is close to random when using the SVM model. However, after FS, the performance improves from 58% to 67%, while maintaining the same SD. The strong results in CV may indicate overfitting, even though the goal of reducing features is to mitigate its risk. So, we can look at the test set performance, to obtain a more objective measure on unseen data.

For the BioVRSea dataset, the test set results are similar to the CV results, which suggests good generalization and provides insight into the sparsity of the results. In contrast, for the SINPAIN dataset, the test performances exceed those of the CV. This could be due to the smaller size of the dataset, which may lead to more variable results across folds (high standard deviation) and less reliable generalization.

Due to their intrinsic differences in model structure and learning mechanisms, RF and SVM can exhibit varying performance on the same data, with RF benefiting from its ensemble approach and ability to handle non-linearities. In contrast, SVM's performance is more sensitive to data distribution and the choice of kernel. These differences highlight that FS remains effective regardless of the chosen algorithm, as RF and SVM perform well on both SINPAIN and BioVRSea datasets. In fact, the results on the BioVRSea dataset (Table 4) are quite comparable between the models, likely due to the complexity of the classification task, where both algorithms can effectively capture relationships between features. However, in the case of COP data, SVM outperforms RF by capturing feature relationships more effectively. Considering the SINPAIN classification results, Table 6, the RF OA classification with muscle features is more effective than SVM. A possible explanation is the fact that the data are not easily separable by the SVM algorithm because of their spatial distribution. This effect is also present when using the complete subset of features, as it includes muscle information.

The aforementioned absent level problem [63] should head towards a decrease in performance when adding the categorical features. The analysis of the BioVRSea dataset shows the impact of categorical features, in RF and SVM classifications. Initially, the introduction of the categorical feature *gender* led to decreased performance in RF [Table 3], which aligns with the aforementioned absent level problem. When a second categorical feature, the lifestyle index, was added, the performance metric did not change when compared to the only *gender* one. The differences observed between the various executions of the algorithm with and without categorical features are quite small, although there were only a few of them. These differences can lead to further investigation, but they fall outside the scope of this discussion. Despite these minor dissimilarities, the overall results are robust and not significantly affected by the presence of categorical features in the dataset.

The goal of the approach is not only to have an effective FS algorithm but also to be as scalable and modular as possible. Indeed, the selected algorithm has been considered the most suitable for our case, as discussed before, but it could not be the same when dealing with different data structures. For example, the merging threshold of 75% is one of the tunable parameters. Future work will include an analysis to explore the impact of different threshold values on models' performance.

Analyzing the BioVRSea performances and reduction, the dataset containing all the features, including the categoricals, is unbalanced with a ratio of 1875 features to 257 subjects, over 7:1. This imbalance can lead to overfitting and stall model performance. However, after

applying FS, the ratio is reduced significantly, to 439 features against the same number of subjects. Moreover, the performances increased in both RF and SVM without impacting the SD in CV; on the contrary, they decreased. Instead, in terms of biosignal, the final feature set composition complies with the original one; and, in this regard, it is worth noting that when feeding the algorithm with features, they are all equals; it has no clue of their origins. The majority of the final set is composed of muscle features, and the performances of the full dataset is similar to the EMG one. However, being the postural control task a difficult one, from these results we cannot conclude that the EMG signals are enough to characterize the phenomenon, indeed the EEG and COP features can classify the effect of the BioVRSea as well, as the results confirm. Moreover, in the final set, they did not show EMG dominance over the other features, for example, by an increase in percentage. Although it is not possible to precisely compare with the results of [41] because of the different metric and biosignal composition, the accuracy using RF is 74.7%, instead in this work RF performance metric after the FS is never reported to be under 77%.

Analyzing the SINPAIN performances and reduction, the dataset containing all the features is unbalanced with a ratio of 1082 features to 102 samples, over 10:1; after the FS, the ratio has decreased to 341 to 102, over 3:1. Also in this case, the performances have increased after the FS with an SVM peak performance on the test set of 94% in Table 5. The SD decreased as well, having more accurate ML models. Again, in terms of biosignals percentage, the final set composition complies with the original one. Although, the final set *F* of the complete dataset, is composed of fewer features from each feature category when compared with the homogeneous FS. For instance, the MRI features extracted when focusing on them are 237, whilst the number of MRI features extracted using the complete one are only 178. This may be because of emerging properties when in the presence of a composition of different features, leading to the need for less data to represent the same information.

The SINPAIN dataset allows us to test the algorithm's capability to extract the best feature set using a limited number of features. The starting number of cartilage features is only 16, and the algorithm selects half of it. Not only does it reduce the number of features, but the selected ones significantly improve the performances of RF and SVM in CV, both in the considered metric and SD. It is unlikely that this behavior is due to overfitting, and the test set's performance confirms this.

In [35] the authors performed a FS by extracting the top 50 features from different models. Considering OA classification on MRI, the reported F1 score in RF in 10-fold CV is  $87.90\% \pm 6.10$  and with SVM  $95.54\% \pm 4.35$ . Their higher score can be because of the tuning of hyperparameters, which fall outside the scope of this work. Indeed, the ensemble FS algorithm is not tuned on any particular dataset, in fact, its objective is to provide an ensemble FS approach. Moreover, the classification models included are only meant to explore whether the reduced features can enhance the ability of AI to distinguish between categories, rather than being already clinically applicable.

In [34] the authors extracted 8 cartilage features from the original 16 and obtained  $0.81\% \pm 0.03$  performance metric with SVM and linear kernel. Our approach also reduces the number of features to 8 reaching  $0.80\% \pm 0.07$  performance metric. The comparable results show that the selected features contain enough information to be able to classify OA comparably to a custom feature extraction method.

#### 4.1. Limitation and strength

This study presents both strengths and limitations, acknowledging challenges that may affect the algorithm's effectiveness and its generalizability.

The selected scaling, the Z-score normalization, allows the preservation of the relationship between the variables, ensuring that every

feature has the same distribution, avoids domination, and is less affected by the presence of outliers [65]. The drawback is that when dealing with non-normal variables, the standard deviation is not a meaningful parameter, and the scaling is ineffective. However, the chosen FS algorithm is scale insensitive, nullifying the scaling drawback. But it is not useless because, in gradient-based algorithms, scaling can improve algorithm convergence speed, especially in high dimensional spaces [66]. One of the algorithm's main characteristics is using different seeds when dealing with random choices, leading to a sort of independence from the seed and allowing perfect replicability. However, they are fixed, and changing them can probably impact the selected features. To overcome this drawback, the number of iterations of the algorithm is set to 50, which has been considered a good tradeoff between time and performance, keeping the goal of generalization [50]. The number of seeds might be considered to the main limitation of the study. In the algorithm presented, the maximum number of features that can theoretically be selected is less than 2500. From FS A, the feature number is capped at 50; these features are then used as a training set for FS B. Considering the extreme case in which FS B is selecting all of them from this iteration, we extract 50 features. This process is repeated for each seed, so 50 times. The extreme case is when each iteration selects a different feature set of 50 features, leading to 2500. The actual number is less than 2500 because of the 75% quartile selection on  $F_R$  and  $F_W$ . However, this threshold does not impact the presented results as the largest input dataset contains only 1875 features. For larger datasets, increasing the number of seeds or the maximum number of selectable features could be considered. The opposite extreme case is when  $F_p$  is a disjoint set from  $F_w$  and  $F_R$ , which is a really rare case because it may be possible only because of different ranking and quartile selection. However, the dataset is composed of biomedical data, and the sets are extracted with the same classification target; in such cases, tuning the parameters and inspecting the sets can give insights into feature importance for further clinical considerations. It is important to notice that the purpose of this algorithm is not to provide a monolithic FS approach but to be scalable to fit to the healthcare data analysis. In the future, it may be worth further exploration of the behavior of the algorithm when changing the number of seeds; intuitively, increasing the number of seeds could lead to a more accurate selection in a bigger multidimensional space.

Scalability is a key property of modern FS algorithms, and the proposed one is flexible on several points. First is the selection algorithm, where users can choose from different FS strategies based on their needs and the nature of the dataset. Additionally, the number of seeds can be tuned for the robustness and diversity of feature subsets selected during the process. Moreover, changing the regularization strategy allows model complexity control and can prevent overfitting. Instead, changing the percentile threshold or merging criteria can lead to different final datasets. The diverse customizations also include changing the maximum number of features allowed and RFE criteria. The results of our feature selection and classification framework have practical value in clinical decision support, as they help reduce the complexity of data analysis while preserving diagnostic performance. This can assist clinicians and researchers in focusing on the most relevant features, leading to more efficient resource allocation, reduced diagnostic time, and potentially more interpretable ML models in real-world healthcare settings.

#### 4.2. Conclusions

The multimetric nature and complex dimensionality of modern healthcare datasets often affect the performances of artificial intelligence tools for clinical evaluation. The proposed ensemble FS method has proved to be able to effectively reduce features in two datasets, postural control BioVRSea and knee OA cartilage study SINPAIN, composed respectively of a large number of features extracted from biosignals and medical images. Moreover, using ML models, the reduced feature sets increase the F1 score without compromising the SD when evaluated in CV, allowing customization to different scenarios and datasets, relying on the algorithm's scalability. These aspects are es-

sential for extracting health-relevant features for better clinical interpretation and development of possible future treatments, indicating its potential as a generalizable tool in healthcare studies.

#### CRedit authorship contribution statement

**Ida Maruotto:** Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Federica Kiyomi Ciliberti:** Writing – original draft, Data curation. **Paolo Gargiulo:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Marco Recenti:** Writing – review & editing, Writing – original draft, Methodology, Supervision, Data curation, Conceptualization.

#### Statements of ethical approval

The BioVRSea studies involving human participants were reviewed and approved by the Icelandic National Bioethics Committee (Application no: 17-183-S1). All the patients/participants provided their written informed consent to participate in the study.

The Sinpain study was approved by the Icelandic Bioethics Commission (approval number: VSN-19-050V1). Prior to any scans, subjects provided comprehensive written consent, ensuring transparency and accountability. To ensure patient privacy, all data are anonymized and securely stored on a dedicated hospital server approved by the Data Protection Unit.

#### Funding

This research was funded by RANNIS, Icelandic Research Fund (Grant of Excellence n: 239612-051) and by European Union (Horizon 2024 - CORDIS Grant agreement ID: 101057778).

#### Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgments

The authors would like to thank all the participants in the BioVRSea and in the Sinpain study. The authors would also like to thank the members of the Institute of Biomedical and Neural Engineering at Reykjavik University who contributed to the BioVRSea acquisitions since 2020.

#### Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.compbiomed.2025.110812>.

#### Data availability

The Feature Selection Software proposed is available open source on the GitHub repository [maruotto/IBNE\\_ScalableFS](https://github.com/maruotto/IBNE_ScalableFS). BioVRSea and SinPain datasets are partially available respectively at <https://biovrsea.is> and <https://restore-project.ru.is>. Both complete datasets are available upon request: please contact Prof. Paolo Gargiulo ([paolo@ru.is](mailto:paolo@ru.is)) or Marco Recenti ([marcor@ru.is](mailto:marcor@ru.is)).

## References

- [1] M. Verleysen, D. François, The curse of dimensionality in data mining and time series prediction, in: *International Work-Conference on Artificial Neural Networks*, Springer, 2005, pp. 758–770.
- [2] V. Berisha, C. Krantsevich, P.R. Hahn, G. Dasarathy, P. Turaga, J. Liss, Digital medicine and the curse of dimensionality, *NPJ Digit. Med.* 4 (1) (2021) 153.
- [3] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, *J. Mach. Learn. Res.* 3 (Mar) (2003) 1157–1182.
- [4] G. Chandrashekar, F. Sahin, A survey on feature selection methods, *Comput. Electr. Eng.* 40 (1) (2014) 16–28, 40th-year commemorative issue, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0045790613003066>.
- [5] N. Sánchez-Maróño, A. Alonso-Betanzos, M. Tombilla-Sanromán, Filter methods for feature selection – A comparative study, in: *Intelligent Data Engineering and Automated Learning - IDEAL 2007*, Springer Berlin Heidelberg, 2007, pp. 178–187, [http://dx.doi.org/10.1007/978-3-540-77226-2\\_19](http://dx.doi.org/10.1007/978-3-540-77226-2_19).
- [6] N. El Aboudi, L. Benhlila, Review on wrapper feature selection approaches, in: *2016 International Conference on Engineering & MIS, ICEMIS*, 2016, pp. 1–5.
- [7] L. Melkumova, S. Shatsikh, Comparing ridge and LASSO estimators for data analysis, *Procedia Eng.* 201 (2017) 746–755, <http://dx.doi.org/10.1016/j.proeng.2017.09.615>.
- [8] I.T. Jolliffe, *Principal Component Analysis for Special Types of Data*, Springer, 2002.
- [9] S.A. Alowais, et al., Revolutionizing healthcare: the role of artificial intelligence in clinical practice, *BMC Med. Educ.* 23 (1) (2023) <http://dx.doi.org/10.1186/s12909-023-04698-z>.
- [10] M. Recenti, Advancing clinical evaluation and diagnostics with artificial intelligence technologies, *Dep. Eng. Reyk. Univ.* (2023).
- [11] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, *Futur. Heal. J.* 6 (2) (2019) 94–98, <http://dx.doi.org/10.7861/futurehosp.6-2-94>.
- [12] D. Park, M. Lee, S.E. Park, J.-K. Seong, I. Youn, Determination of optimal heart rate variability features based on SVM-recursive feature elimination for cumulative stress monitoring using ECG sensor, *Sensors* 18 (7) (2018) 2387, <http://dx.doi.org/10.3390/s18072387>.
- [13] H. Li, D. Yuan, X. Ma, D. Cui, L. Cao, Genetic algorithm for the optimization of features and neural networks in ECG signals classification, *Sci. Rep.* 7 (1) (2017) <http://dx.doi.org/10.1038/srep41011>.
- [14] K.K. Patro, A. Jaya Prakash, M. Jayamanmadha Rao, P. Rajesh Kumar, An efficient optimized feature selection with machine learning approach for ECG biometric recognition, *IETE J. Res.* 68 (4) (2020) 2743–2754, <http://dx.doi.org/10.1080/03772063.2020.1725663>.
- [15] M.H. Bhatti, J. Khan, M.U.G. Khan, R. Iqbal, M. Aloqaily, Y. Jararweh, B. Gupta, Soft computing-based EEG classification by optimal feature selection and neural networks, *IEEE Trans. Ind. Inform.* 15 (10) (2019) 5747–5754, <http://dx.doi.org/10.1109/tii.2019.2925624>.
- [16] L. Duan, H. Ge, W. Ma, J. Miao, EEG feature selection method based on decision tree, in: F. Liu, D.-H. Lee, R. Lagoa, S. Kumar (Eds.), *Bio-Med. Mater. Eng.* 26 (s1) (2015) S1019–S1025, <http://dx.doi.org/10.3233/bme-151397>.
- [17] X. Li, B. Hu, S. Sun, H. Cai, EEG-based mild depressive detection using feature selection methods and classifiers, *Comput. Methods Programs Biomed.* 136 (2016) 151–161, <http://dx.doi.org/10.1016/j.cmpb.2016.08.010>.
- [18] Y. Xue, X. Han, Z. Wang, Self-adaptive weight based on dual-attention for differentiable neural architecture search, *IEEE Trans. Ind. Inform.* 20 (4) (2024) 6394–6403.
- [19] Y. Xue, K. Chen, F. Neri, Differentiable architecture search with attention mechanisms for generative adversarial networks, *IEEE Trans. Emerg. Top. Comput. Intell.* 8 (4) (2024) 3141–3151.
- [20] P. Sahu, B.K. Singh, N. Nirala, An improved feature selection approach using global best guided Gaussian artificial bee colony for EMG classification, *Biomed. Signal Process. Control.* 80 (2023) 104399, <http://dx.doi.org/10.1016/j.bspc.2022.104399>.
- [21] J. Too, A.R. Abdullah, N. Mohd Saad, W. Tee, EMG feature selection and classification using a pbest-guide binary particle swarm optimization, *Computation* 7 (1) (2019) 12, <http://dx.doi.org/10.3390/computation7010012>.
- [22] R. Cheng, Y. Jin, A competitive swarm optimizer for large scale optimization, *IEEE Trans. Cybern.* 45 (2) (2015) 191–204.
- [23] Y. Hu, Y. Zhang, D. Gong, Multiobjective particle swarm optimization for feature selection with fuzzy cost, *IEEE Trans. Cybern.* 51 (2) (2021) 874–888.
- [24] Y. Zhang, D. wei Gong, X. zhi Gao, T. Tian, X. yan Sun, Binary differential evolution with self-learning for multi-objective feature selection, *Inform. Sci.* 507 (2020) 67–85, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0020025519307819>.
- [25] B. Xue, M. Zhang, W.N. Browne, Particle swarm optimization for feature selection in classification: A multi-objective approach, *IEEE Trans. Cybern.* 43 (6) (2013) 1656–1671.
- [26] R. Alzubi, N. Ramzan, H. Alzoubi, A. Amira, A hybrid feature selection method for complex diseases SNPs, *IEEE Access* 6 (2018) 1292–1301.
- [27] P. V, J. V, Hybrid feature selection technique for prediction of cardiovascular diseases, *Mater. Today: Proc.* 81 (2023) 336–340, <http://dx.doi.org/10.1016/j.matpr.2021.03.225>.
- [28] Y. Xue, C. Zhang, A novel importance-guided particle swarm optimization based on MLP for solving large-scale feature selection problems, *Swarm Evol. Comput.* 91 (2024) 101760, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2210650224002980>.
- [29] Y. Xue, C. Zhang, F. Neri, M. Gabbouj, Y. Zhang, An external attention-based feature ranker for large-scale feature selection, *Knowl.-Based Syst.* 281 (2023) 111084, <http://dx.doi.org/10.1016/j.knsys.2023.111084>.
- [30] N. Naheed, M. Shaheen, S. Ali Khan, M. Alawairdhi, M. Attique Khan, Importance of features selection, attributes selection, challenges and future directions for medical imaging data: A review, *Comput. Model. Eng. Sci.* 125 (1) (2020) 315–344, <http://dx.doi.org/10.32604/cmescs.2020.011380>.
- [31] S. Ibrahim, S. Nazir, S.A. Velastin, Feature selection using correlation analysis and principal component analysis for accurate breast cancer diagnosis, *J. Imaging* 7 (11) (2021) 225, <http://dx.doi.org/10.3390/jimaging7110225>.
- [32] I. Reda, A. Khalil, M. Elmogy, A. Abou El-Fetouh, A. Shalaby, M. Abou El-Ghar, A. Elmaghraby, M. Ghazal, A. El-Baz, Deep learning role in early diagnosis of prostate cancer, *Technol. Cancer Res. Treat.* 17 (2018) 153303461877553, <http://dx.doi.org/10.1177/1533034618775530>.
- [33] F.K. Ciliberti, L. Guerrini, A.E. Gunnarsson, M. Recenti, D. Jacob, V. Cangiano, Y.A. Tesfahunegn, A.S. Islind, F. Tortorella, M. Tsirilaki, H. Jónsson, P. Gargiulo, R. Aubonnet, CT- and MRI-based 3D reconstruction of knee joint to assess cartilage and bone, *Diagnostics* 12 (2) (2022) [Online]. Available: <https://www.mdpi.com/2075-4418/12/2/279>.
- [34] F.K. Ciliberti, I. Maruotto, H.J. Jónsson, M.K. Gíslason, P. Gargiulo, Multimetric evaluation of knee cartilage degeneration, in: *Proceedings of IEEE MetroXRaine 2024*, 2024.
- [35] F. Angelone, F.K. Ciliberti, G.P. Tobia, H. Jónsson, Jr., A.M. Ponsiglione, M.K. Gíslason, F. Tortorella, F. Amato, P. Gargiulo, Innovative diagnostic approaches for predicting knee cartilage degeneration in osteoarthritis patients: A radiomics-based study, *Inf. Syst. Front.* (2024) 1–23.
- [36] D. Jacob, R. Aubonnet, M. Recenti, S.A. Audardottir, T. Ida Ivarsdottir, B. Burgunder, I.M.I. Escalona, A. Colacino, A. Bjornsdottir, H. Petersen, P. Gargiulo, Assessing early-stage parkinson's disease using BioVRSea, in: *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering, MetroXRaine*, 2022, pp. 271–276.
- [37] D. Jacob, L. Guerrini, F. Pescaglia, S. Pierucci, C. Gelormini, V. Minutolo, A. Frattini, G. Di Lorenzo, H. Petersen, P. Gargiulo, Adaptation strategies and neurophysiological response in early-stage parkinson's disease: BioVRSea approach, *Front. Hum. Neurosci.* 17 (2023) <http://dx.doi.org/10.3389/fnhum.2023.1197142>.
- [38] M. Recenti, D. Jacob, R. Aubonnet, B. Burgunder, I.M.I. Escalona, A.E. Gunnarsson, F.K. Ciliberti, R. Forni, L. Donisi, H. Petersen, P. Gargiulo, Predicting lifestyle using BioVRSea multi-biometric paradigms, in: *2022 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering, MetroXRaine*, 2022, pp. 329–334.
- [39] S.A. Stehle, R. Aubonnet, M. Hassan, M. Recenti, D. Jacob, H. Petersen, P. Gargiulo, Predicting postural control adaptation measuring EEG, EMG, and center of pressure changes: BioVRSea paradigm, *Front. Hum. Neurosci.* 16 (2022) <http://dx.doi.org/10.3389/fnhum.2022.1038976>.
- [40] D. Jacob, I.S. Unnsteinsdóttir Kristensen, R. Aubonnet, M. Recenti, L. Donisi, C. Ricciardi, et al., Towards defining biomarkers to evaluate concussions using virtual reality and a moving platform (BioVRSea), *Sci. Rep.* 12 (1) (2022) <http://dx.doi.org/10.1038/s41598-022-12822-0>.
- [41] M. Recenti, L. Guerrini, A. Lindemann, S. Pierucci, C. Ricciardi, A.M. Ponsiglione, H. Petersen, P. Gargiulo, Heart rate variability during a complex postural control task with the BioVRSea paradigm, in: *2023 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering, MetroXRaine*, 2023, pp. 876–881.
- [42] M. Recenti, F. Pescaglia, L. Guerrini, I. Maruotto, C. Gelormini, D. Jacob, R. Aubonnet, H. Petersen, P. Gargiulo, New frontiers in postural control and motion sickness assessment: The BioVRSea paradigm, in: *2024 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering, MetroXRaine*, 2024, pp. 1034–1039.
- [43] M. Recenti, C. Ricciardi, R. Aubonnet, I. Picone, D. Jacob, et al., Toward predicting motion sickness using virtual reality and a moving platform assessing brain, muscles, and heart signals, *Front. Bioeng. Biotechnol.* 9 (2021) <http://dx.doi.org/10.3389/fbioe.2021.635661>.
- [44] J.F. Golding, Predicting individual differences in motion sickness susceptibility by questionnaire, *Pers. Individ. Differ.* 41 (2) (2006) 237–248.
- [45] R. Aubonnet, J. Ramos, M. Recenti, D. Jacob, F. Ciliberti, L. Guerrini, M.K. Gíslason, O. Sigurjonsson, M. Tsirilaki, J. Halldór Jónsson, P. Gargiulo, Toward new assessment of knee cartilage degeneration, *Cartilage* 14 (3) (2023) 351–374, <http://dx.doi.org/10.1177/19476035221144746>, PMID: 36541701.
- [46] J.J. Van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R.G. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, H.J. Aerts, Computational radiomics system to decode the radiographic phenotype, *Cancer Res.* 77 (21) (2017) e104–e107.

- [47] F. Angelone, F.K. Ciliberti, H.j. Jónsson, M.K. Gíslason, P. Gargiulo, Knee cartilage degradation in the medial and lateral anatomical compartments: A radiomics study, in: *Proceedings of IEEE MetroXRaine 2024*, 2024.
- [48] F.K. Ciliberti, G. Pianegonda, H.j. Jónsson, M.K. Gíslason, P. Gargiulo, How muscle and intramuscular adipose tissue volume and density progress in knee osteoarthritis-affected subjects: a virtual histology study, in: *Proceedings of IEEE MetroXRaine 2024*, 2024.
- [49] R. Forni, C. Gelormini, C. Corsi, P. Gargiulo, Virtual histology of the heart through CT imaging: Preliminary results of a novel noninvasive approach for cardiac tissue characterization, in: *2023 Computing in Cardiology (CinC)*, vol. 50, 2023, pp. 1–4.
- [50] C. Colas, O. Sigaud, P.-Y. Oudeyer, How many random seeds? Statistical power analysis in deep reinforcement learning experiments, 2018, *ArXiv abs/1806.08295*, [Online]. Available: <https://api.semanticscholar.org/CorpusID:49350736>.
- [51] A.K. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, *Pattern Recognit.* 38 (2005) 2270–2285.
- [52] J. Singh, S. Bagga, R. Kaur, Software-based prediction of liver disease with feature selection and classification techniques, *Procedia Comput. Sci.* 167 (2020) 1970–1980.
- [53] J.H. Friedman, Stochastic gradient boosting, *Comput. Statist. Data Anal.* 38 (4) (2002) 367–378.
- [54] P. Theerthagiri, Predictive analysis of cardiovascular disease using gradient boosting based learning and recursive feature elimination technique, *Intell. Syst. Appl.* 16 (2022) 200121, [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S266730532200059X>.
- [55] J.P. Barddal, F. Enembreck, Learning regularized hoeffding trees from data streams, in: *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, 2019.
- [56] R. Tibshirani, The lasso problem and uniqueness, *Electron. J. Stat.* 7 (2012) 1456–1490.
- [57] B.F. Darst, K.C. Malecki, C.D. Engelman, Using recursive feature elimination in random forest to account for correlated variables in high dimensional data, *BMC Genet.* 19 (2018) 1–6.
- [58] M. Črepinšek, S.-H. Liu, M. Mernik, Exploration and exploitation in evolutionary algorithms: A survey, *ACM Comput. Surv.* 45 (3) (2013) 1–33.
- [59] I. De Falco, U. Scafuri, E. Tarantino, A. Della Cioppa, A distributed differential evolution approach for mapping in a grid environment, in: *15th EUROMICRO International Conference on Parallel, Distributed and Network-Based Processing, PDP'07*, 2007, pp. 442–449.
- [60] L. Breiman, Random forests, *Mach. Learn.* 45 (2001) 5–32.
- [61] S. Suthaharan, S. Suthaharan, Support vector machine, in: *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, Springer, 2016, pp. 207–235.
- [62] S. Panja, A. Chatterjee, G. Yasmin, Kernel functions of SVM: A comparison and optimal solution, in: A.K. Luhach, D. Singh, P.-A. Hsiung, K.B.G. Hawari, P. Lingras, P.K. Singh (Eds.), *Advanced Informatics for Computing Research*, 2019.
- [63] T.C. Au, Random forests, decision trees, and categorical predictors: The “absent levels” problem, *J. Mach. Learn. Res.* 19 (45) (2018) 1–30, [Online]. Available: <http://jmlr.org/papers/v19/16-474.html>.
- [64] S.A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M.A. Riegler, P. Halvorsen, S. Parasa, On evaluation metrics for medical applications of artificial intelligence, *Sci. Rep.* 12 (1) (2022) <http://dx.doi.org/10.1038/s41598-022-09954-8>.
- [65] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, *Appl. Soft Comput.* 97 (2020) 105524.
- [66] Z. Chen, C. Xu, Global convergence of a general sampling algorithm for dynamic nonlinear stochastic programs, *Numer. Funct. Anal. Optim.* 23 (2002) 495–514.