Computers in Biology and Medicine 180 (2024) 108999

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Positive-Unlabelled learning for identifying new candidate Dietary Restriction-related genes among ageing-related genes

Jorge Paz-Ruza [a,*], Alex A. Freitas [b], Amparo Alonso-Betanzos [a], Bertha Guijarro-Berdiñas [a]

[a] LIDIA Group, CITIC, Universidade da Coruña, Campus de Elviña s/n, A Coruña 15071, Spain
[b] School of Computing, University of Kent, Canterbury CT2 7FS, United Kingdom

ARTICLE INFO

*Keywords:*
Machine Learning
Positive-Unlabelled learning
Bioinformatics
Ageing
Dietary Restriction

ABSTRACT

Dietary Restriction (DR) is one of the most popular anti-ageing interventions; recently, Machine Learning (ML) has been explored to identify potential DR-related genes among ageing-related genes, aiming to minimize costly wet lab experiments needed to expand our knowledge on DR. However, to train a model from positive (DR-related) and negative (non-DR-related) examples, the existing ML approach naively labels genes without known DR relation as negative examples, assuming that lack of DR-related annotation for a gene represents evidence of absence of DR-relatedness, rather than absence of evidence. This hinders the reliability of the negative examples (non-DR-related genes) and the method's ability to identify novel DR-related genes. This work introduces a novel gene prioritization method based on the two-step Positive-Unlabelled (PU) Learning paradigm: using a similarity-based, KNN-inspired approach, our method first selects reliable negative examples among the genes without known DR associations. Then, these reliable negatives and all known positives are used to train a classifier that effectively differentiates DR-related and non-DR-related genes, which is finally employed to generate a more reliable ranking of promising genes for novel DR-relatedness. Our method significantly outperforms ($p < 0.05$) the existing state-of-the-art approach in three predictive accuracy metrics with up to $\sim 40\%$ lower computational cost in the best case, and we identify 4 new promising DR-related genes (PRKAB1, PRKAB2, IRS2, PRKAG1), all with evidence from the existing literature supporting their potential DR-related role.

## 1. Introduction

Ageing is a biological process characterized by a progressive decline in physiological function and increased susceptibility to age-related diseases. As it is a complex phenomenon influenced by both genetic and environmental factors [1], understanding the genetic basis of ageing is crucial for deciphering its underlying mechanisms and developing methods to promote healthy ageing. The scientific community has invested a great amount of research efforts into understanding the biological processes involved in ageing; particularly, genetic studies have identified numerous genes and pathways associated with ageing, offering insights into potential targets for anti-ageing therapeutic interventions [2–4].

One of the most promising and studied approaches to extend lifespan and delay the onset of age-related diseases is Dietary Restriction (DR), which involves reducing nutrient intake (and typically, calorie intake) without causing malnutrition [5] and has been shown to extend lifespan and improve long-term health in various model organisms [6]. By modulating various cellular pathways, such as insulin signalling, sirtuin activation, or autophagy induction [7–9], DR promotes cellular stress resistance and metabolic efficiency, reducing the risk of age-related pathologies such as cardiovascular disease, cancer, and neurodegeneration [10,11].

The research efforts in ageing and other biomedical areas greatly increased the magnitude and complexity of available biological data; as a solution, Machine Learning (ML) has emerged as a powerful tool to facilitate the analysis of large-scale biological datasets and uncover hidden patterns [12,13]. ML techniques have been widely applied in ageing-related research, including the prediction of lifespan of model organisms, the identification of molecular signatures of ageing, and the association of metabolic pathways with ageing-related diseases [14,15].

In the particular topic of DR, Magdaleno et al. [16] recently employed ML to classify ageing-related genes into DR-related and non-DR-related genes, in order to identify candidate DR-related genes among

* Corresponding author.
*E-mail addresses:* j.ruza@udc.es (J. Paz-Ruza), A.A.Freitas@kent.ac.uk (A.A. Freitas), amparo.alonso.betanzos@udc.es (A. Alonso-Betanzos), berta.guijarro@udc.es (B. Guijarro-Berdiñas).

https://doi.org/10.1016/j.compbiomed.2024.108999
Received 14 June 2024; Received in revised form 25 July 2024; Accepted 5 August 2024

ageing-related genes not currently annotated as DR-related. Utilizing various biological features, such as pathway information from PathDIP [17], Gene Ontology (GO) terms [18], KEGG pathways [19], or coexpression data, they trained decision tree-based ensemble classifiers under a binary classification task to ultimately produce a ranking of promising DR-related genes for wet-lab verification.

However, the pipeline described in [16] holds a significant limitation: to provide training examples with binary labels to the classifier, Magdaleno et al. assumed that all ageing-related genes without experimental evidence of DR-relatedness can be considered as non-DR-related, i.e. labelled as negative training examples. As the authors acknowledge, an absence of evidence does not equate to evidence of absence of DR-relatedness, meaning the classifier was trained with an unknown amount of incorrectly labelled negative examples, hindering its learning and therefore its ability to correctly identify new DR-related genes.

This type of data, commonly referred to as Positive-Unlabelled (PU) data, corresponds to cases where a subset of the examples are labelled as known positives and, while the rest is unlabelled and comprises both negative and positive examples [20]. This is common in bioinformatics due to the cost of obtaining annotations through wet-lab experimentation [21], but ignoring unlabelled data or treating it as negative (as usual in the literature) leads to biased and suboptimal models in many ML tasks [22]. As a solution, PU Learning is an ML paradigm specifically designed for improving the quality and predictive power of classifiers in settings that involve PU data, acknowledging unlabelled examples as such throughout the ML pipeline [23].

In this work, we propose a PU Learning method to enhance the prediction of novel candidate DR-related genes among ageing-related genes. Our approach addresses the limitations of the Magdaleno et al.'s [16] existing methodology by properly accounting for the unlabelled data and exploiting its potential to improve the quality of predictions. Specifically, we propose a similarity-based two-step PU Learning strategy to improve the training process and the predictive power of classifiers. The proposed method has surpassed the performance of the state-of-the-art non-PU Learning method in our target task of predicting novel DR-related genes among ageing-related genes. In addition, we use this PU learning methodology to generate a more reliable list of top candidate genes for novel DR-relatedness supported by evidence in the literature of the field.

We underline three main contributions of this work:

- Firstly, the design, implementation and experimental validation of a PU learning method for gene prioritization, and its application to improve the identification of DR-related genes among ageing-related genes. In this task, our method significantly outperformed the existing state-of-the-art (non-PU) method in all three used predictive performance metrics, on real-world ageing-related gene datasets, while reducing the required computational overhead to train and employ the model for DR-related gene identification.
- Secondly, the novel use of the proposed PU Learning method to produce a more reliable list of top candidate DR-related genes, compared to Magdaleno et al.'s [16] state-of-the-art approach, owing to our method's superior performance in computational experiments.
- Lastly, a curation of the relevant literature, which identified evidence supporting the potential DR-relatedness of our method's top candidate genes, further motivating wet-lab experiments that could validate the predicted DR-relatedness of the proposed genes.

The remainder of this paper is structured as follows. In Section 2 we discuss our target task and existing methodology for identifying new DR-related genes, highlight its limitations, and introduce basic notions of the PU Learning ML paradigm and the need for it in the task. Section 3 presents our proposed PU Learning method, detailing how it enhances the training process to improve classifiers' power. In Section 4, we describe the experimental setup, including features, classifiers and evaluation metrics. Section 5 presents the results of our experiments, comparing our PU Learning approach against the state-of-the-art non-PU approach. Section 6 concludes the paper and discusses avenues for future research.

## 2. Background

In this section we formalize the task of interest, cover the existing ML-based approach used to solve it and its limitations, and introduce basic concepts of the PU Learning paradigm in ML and its need for training classifiers in this task.

### 2.1. Task formulation

The task at hand is to find new candidate genes related to DR among ageing-related genes. Formally, let $\mathcal{G}_{\text{AGE}}$ be the set of known ageing-related genes, each associated with a vector of $n$ biological features $\mathbf{x}_g = (f_1, f_2, \ldots, f_n)$. We assume there exists a subset of genes $\mathcal{G}_{\text{DR}\cap\text{AGE}^+}$ involved in DR used as anti-ageing intervention. Among these, a smaller subset of genes $\mathcal{G}_{\text{DR}\cap\text{AGE}}$ have experimental evidence of DR-relatedness. Reasonably, not all ageing-related genes have a relation to DR, and not all DR-related genes have been identified experimentally, and therefore:

$$\mathcal{G}_{\text{DR}\cap\text{AGE}} \subset \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \subset \mathcal{G}_{\text{AGE}} \tag{1}$$

The objective is to identify ageing-related genes $g^*$ without known evidence of DR-relatedness but with high probability of actually being DR-related, such that:

$$g^* = \text{argmax}_{g\in\mathcal{G}_{\text{AGE}}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}}} Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) \tag{2}$$

where $g^*$ are the genes most likely to be found to be DR-related in future wet-lab experiments, constituting a task of knowledge discovery, i.e. finding the subset $\mathcal{G}_{\text{DR}\cap\text{AGE}^+} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}}$.

A general solution of this task is to find a model $\boldsymbol{\Phi} : \mathcal{G}_{\text{AGE}} \rightarrow [0, 1]$ that assigns a score $\boldsymbol{\Phi}(g)$ to each gene $g$ as a measure of its DR-relatedness. Since the only known information is which genes already belong to $\mathcal{G}_{\text{DR}\cap\text{AGE}}$, where trivially $Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) = 1$, the task is surrogated to a binary classification problem: the positive class are those genes that belong to $\mathcal{G}_{\text{DR}\cap\text{AGE}^+}$, and the negative class consists of the remaining ageing-related genes ($\mathcal{G}_{\text{AGE}} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}^+}$).

An important detail of this formulation is that the real positive class is not composed solely of genes with known association with DR, but also with undiscovered DR-relatedness, i.e. that belong to the unknown set $\mathcal{G}_{\text{DR}\cap\text{AGE}^+} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}}$. Similarly, the actual negative class is not the set of genes without known DR-relatedness ($\mathcal{G}_{\text{AGE}} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}}$), but the set of genes without real DR-relatedness ($\mathcal{G}_{\text{AGE}} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}^+}$).

The model $\boldsymbol{\Phi}$ to be designed is then a binary classifier that, given a gene's features, outputs the probability of it belonging to the positive class, such that:

$$\boldsymbol{\Phi}(g) = \begin{cases} x \in [0.5, 1] & g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \\ x \in [0, 0.5) & g \in \mathcal{G}_{\text{AGE}} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \end{cases} \tag{3}$$

Since $\boldsymbol{\Phi}(g)$ can also be interpreted as a probability of DR-relatedness, it approximates the original goal, as seen in Eq. (4): the genes without known DR-relatedness given highest scores by $\boldsymbol{\Phi}$ should be the most promising candidates for DR-relatedness.

$$\begin{aligned} &\text{argmax}_{g\in\mathcal{G}_{\text{AGE}}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}}} \boldsymbol{\Phi}(g) \approx \text{argmax}_{g\in\mathcal{G}_{\text{AGE}}\setminus\mathcal{G}_{\text{DR}\cap\text{AGE}}} Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) \\ &\approx \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}} \end{aligned} \tag{4}$$

Fig. 1 summarizes this discovery task as two steps: (1) surrogating the task to a binary classification to achieve a model which identifies DR-related genes among a set of ageing-related genes, and (2) using the model's predictions on ageing-related genes without known DR-relatedness to generate a ranking of promising, undiscovered DR-related genes.

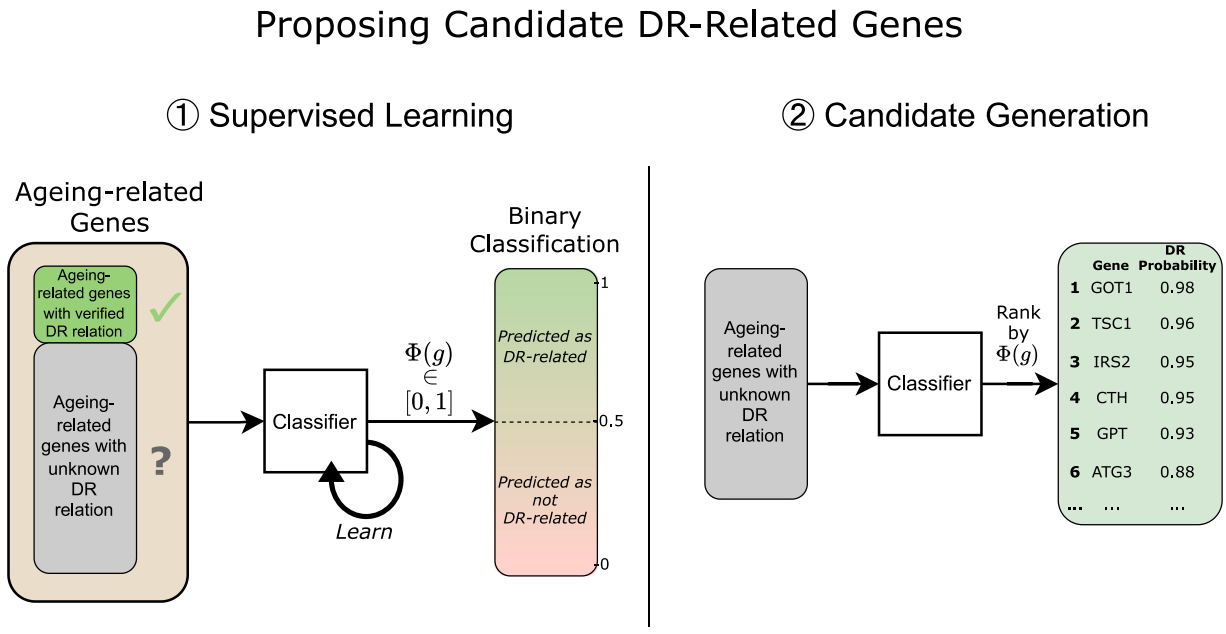


**Fig. 1.** General overview of the two-step modelling to solve the task for proposing potential novel DR-related genes among ageing-related genes.

### 2.2. Existing methodology for DR-related gene identification

Recently, Magdaleno et al. [16] explored the task of predicting novel DR-relatedness among ageing-related genes using ML techniques, showing promising results. By the time of our research, Magdaleno et al.'s work is the only work for identifying novel DR-related genes among ageing-related genes. In addition they used state-of-the-art classifiers for tabular data (decision tree ensembles) [24]. Hence, their approach is the state-of-the-art for this task.

To characterize each gene, they considered a variety of biological features: PathDIP gene-pathway interactions [17], KEGG pertinence [19] and influence [25] descriptors, protein-protein interaction (PPI) [26] adjacency and graph metrics, hierarchical Gene Ontology (GO) terms [18], expression information in tissues from GTEx [27], gene co-expression data [28], and protein descriptors [29]. These features were analysed individually and collectively as potential predictors for DR-relatedness of ageing-related genes.

In terms of classification algorithms, Magdaleno et al. used standard decision tree-based ensemble algorithms for their state-of-the-art performance in tabular data [24], experimenting with Balanced Random Forest (BRF) [30], XGBoost [31], Easy Ensemble Classifier [32], and CatBoost [33], and employing a nested cross-validation to evaluate all combinations of features and classifiers. Interestingly, their results indicated that combining different types of biological features did not necessarily improve performance.

Two best-performing combinations of a feature type and a classifier were identified ({PathDIP, CatBoost} and {GO, BRF}), and then used to predict the DR-relatedness of genes without known DR association and propose a ranking of the most promising candidates.

A fundamental element of the methodology used by Magdaleno et al. [16] lies in the assumption that all genes without known DR relation can be considered negative examples during training. This is, the model is taught to predict as

$$\Phi(g) = 0 \quad \forall g \in \mathcal{G}_{\text{AGE}} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}} \tag{5}$$

but, by the formulation from Section 2.1, this is equivalent to teaching the model that

$$\begin{aligned} Pr(g \in \mathcal{G}_{\text{DR}\cap\text{AGE}^+}) = 0 \quad \forall g \in \mathcal{G}_{\text{AGE}} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}} \\ \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}} = \emptyset \end{aligned} \tag{6}$$

which defeats the purpose of the discovery task: retrieving with the ML algorithm the undiscovered DR-related genes that exist in the set of genes without known DR-relatedness, i.e. finding $\mathcal{G}_{\text{DR}\cap\text{AGE}^+} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}}$. Ultimately, this can lead to a lower performance of the ML model and, in consequence, a less reliable set of candidate genes for DR-relatedness.

In this work, we show the design and usage of a more sophisticated labelling algorithm of training examples based on PU Learning, making it aware that unlabelled examples are not necessarily negative (i.e. that some genes without known DR-relatedness may actually be DR-related), can overcome the aforementioned limitation of Magdaleno et al.'s state-of-the-art approach and improve the identification of novel DR-related genes without adding computational overhead (by comparison with Magdaleno et al.'s approach).

### 2.3. Essential notions of PU Learning

In the context of semi-supervised learning, PU Learning is an ML paradigm designed for scenarios where, rather than the classic positive and negative examples, a dataset is composed of a subset of positive examples $\mathcal{P}$ and a set of unlabelled examples $\mathcal{U}$, which is assumed to contain both positive and negative examples [20]. This paradigm suitable when positive instances are a priority but labelling many instances is impractical or very expensive, such as in gene prioritization or other bioinformatics tasks.

In bioinformatics, PU Learning has been explored in varied tasks [21]. Zheng et al. [34] employed PU Learning with drug-drug interaction data to improve the identification of dangerous adverse reactions in patients with multiple medications. Lan et al. [35] tackled the discovery of drug-target pairs using PU Learning. Kılıç and Mehmet [36] explored PU Learning to derive knowledge on protein-protein interaction networks, and Song et al. [37] used PU Learning to predict sequence-function relationships in large-scale proteomics data. Most recently, PU Learning has shown promising results in discovering toxin-degrading enzymes [38] or predicting the secreted proteins in human body fluids for biomarker identification of diseases [39].

Learning from PU data is not trivial and PU Learning encompasses different strategies, such as biased learning [40,41], incorporation of class prior knowledge [20], or the so-called *two-step* methods [23]; the latter are the focus of this work.

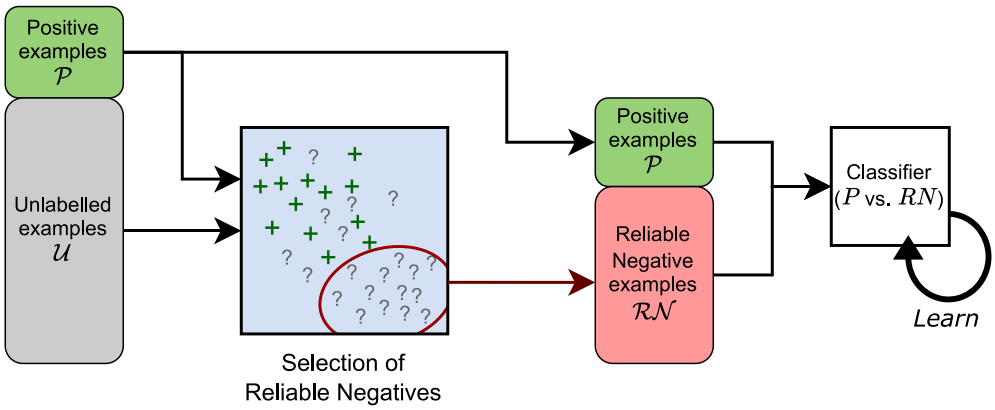


**Fig. 2.** High-level structure of a two-step PU Learning technique.

Two-step methods are based around a core idea: training a binary classifier under the assumption that all unlabelled examples can be considered negative introduces label noise, hinders performance and breaks assumptions of PU tasks; this is a limitation in Magdaleno et al.'s work [16], as discussed in Section 2.2. Alternatively, two-step methods propose training a binary classifier using the positive examples $\mathcal{P}$ and a subset $\mathcal{RN} \subset \mathcal{U}$ of "reliable negatives" extracted from the set of unlabelled examples. If the reliable negatives are correctly identified, even if scarce in number, a binary classifier can be trained with minimal label noise and respecting the PU tasks assumptions, increasing predictive performance. As such, PU learning is of particular interest in bioinformatics and biomedicine [21].

Fig. 2 depicts the classic template for two-step PU Learning methods, with the Unlabelled set $\mathcal{U}$ being distilled into a smaller set of Reliable Negatives $\mathcal{RN}$. Although the number of training examples is reduced in the process, the quality of training data greatly increases by minimizing label noise; ultimately, this improves the efficacy and efficiency of training and obtains classifiers with significantly higher predictive performance in PU data scenarios [20].

## 3. The proposed PU Learning method

This section details the PU Learning method designed to improve the discovery of DR-related genes among ageing-related genes. Instead of treating all unlabelled examples (genes without known relation to DR) as negative examples in training, breaking the assumptions of the knowledge discovery task, we propose a training strategy where the model learns from a refined training set of positive and reliable negative examples extracted from the set of unlabelled genes.

Our PU Learning methodology is a similarity-based method with labelling criteria inspired by a classic nearest neighbour classification (KNN) [42], and can be categorized as a two-step, prior-free method in Bekker and Davis' PU Learning taxonomy [23]. As other methods in this category, we assume ageing-related data respects two assumptions: smoothness of the positive class (DR-related genes exhibit similarities in their biological features) and separability (non-DR-related genes differ from DR-related ones in their biological features). We also assume *Selected Completely at Random* (SCAR) as the underlying labelling mechanism of the positive class: the set of labelled positive examples $\mathcal{G}_{\text{DR}\cap\text{AGE}}$ is an i.i.d. sample of the set of all positive examples $\mathcal{G}_{\text{DR}\cap\text{AGE}^+}$; this is a popular and reasonable assumption in bioinformatics tasks [43].

Given a dataset $\mathcal{D}$ composed of a set $\mathcal{P}$ of genes with known DR relatedness (positive examples) and a set $\mathcal{U}$ of genes without known DR relatedness (unlabelled examples), where each gene is represented with a set of biological features $\mathcal{F}$, a set $\mathcal{RN}$ of genes unlikely to be DR-related (reliable negatives) is obtained as follows:

1. For each pair of genes $(g_i, g_j)$, a similarity metric is computed using their biological feature vectors $x_g = (f_1, \ldots, f_{|\mathcal{F}|})$. Due to the high degree of feature sparsity of the datasets involved and the fact that the datasets have only binary features (see Section 4.1), we choose to use the Jaccard Measure $J(x_{g_i}, x_{g_j})$ [44,45] over other options such as the Euclidean or cosine distances. Specifically, the Jaccard similarity between two ageing-related genes represented by vectors of binary biological features is computed as:

$$J(x_{g_i}, x_{g_j}) = \frac{\sum_{k=1}^{|\mathcal{F}|} x_{g_i}[k] * x_{g_j}[k]}{\sum_{k=1}^{|\mathcal{F}|} x_{g_i}[k] + \sum_{k=1}^{|\mathcal{F}|} x_{g_j}[k] - \sum_{k=1}^{|\mathcal{F}|} x_{g_i}[k] * x_{g_j}[k]} \quad (7)$$

   where the numerator counts features with value 1 in both genes' feature vectors and the denominator counts features with value 1 in only one of the two genes' feature vectors.
2. The set of reliable negatives $\mathcal{RN}$ is initialized with no elements.
3. For each unlabelled gene $g_i$ in the training set:
   (a) Find the $k$ training gene closest to $g_i$ (its $k$ nearest neighbours) based on their pairwise Jaccard similarities.
   (b) Two conditions are checked:
      - Whether the single closest gene to $g_i$ is unlabelled (no known DR relation).
      - Whether the proportion of genes without known relation to DR among the top $k$ nearest neighbours of $g_i$ is higher than a set threshold $t$.
   (c) If these two conditions are met, the gene $g_i$ is added to the set of reliable negatives.

Algorithm 1 contains the logic and Fig. 3 a visual depiction of the process above, constituting the first step of the two-step PU Learning method. The result is a set of reliable negatives $\mathcal{RN}$ with very high confidence of not being DR-related. Formally, $\mathcal{RN}$ should verify:

$$\begin{aligned} &\mathcal{RN} \subset \mathcal{G}_{\text{AGE}} \setminus \mathcal{G}_{\text{DR}\cap\text{AGE}^+} \\ &\mathcal{RN} \cap \mathcal{G}_{\text{DR}\cap\text{AGE}^+} = \emptyset \end{aligned} \quad (8)$$

therefore respecting the assumptions of the discovery task, as laid out in Fig. 1 and unlike the naive approach of treating all unlabelled examples as negatives.

At the second and final step, $\mathcal{P}$ and $\mathcal{RN}$ are used to train the final model $\Phi$ with minimal label noise, improving data quality and the effectiveness and efficiency of the model learning.

This PU Learning method can be easily integrated with any classifier to predict the DR-relatedness of genes without known DR-relation, as this method is classifier-agnostic. Algorithm 2 shows the integration of our PU Learning method inside a standard nested cross-validation

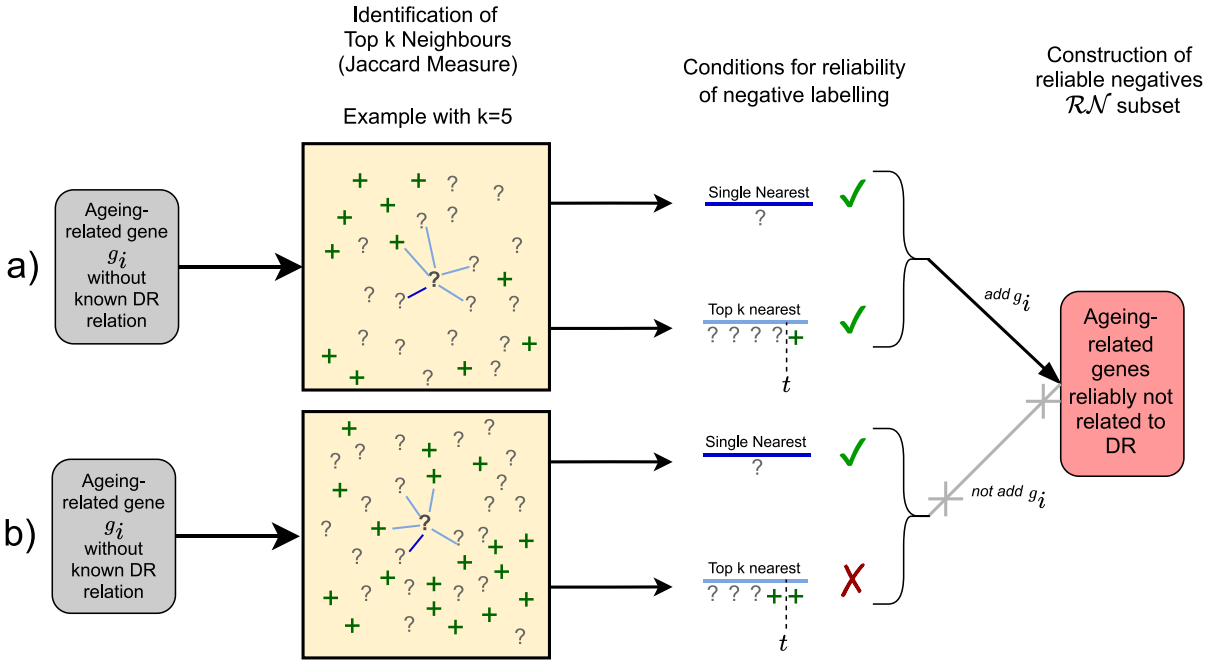


**Fig. 3.** Similarity-based reliable negative selection of the proposed PU Learning algorithm. The threshold $t$ is the minimum proportion of unlabelled examples among the $k$ nearest neighbours of an unlabelled example required to consider it a reliable negative ($k$ and $t$ are tunable hyperparameters; in this example, $k = 5$ and $t = 0.8$). Two different cases are shown: in case (a) the two conditions for a reliable negative are met, i.e. the gene's nearest neighbour and >80% of its $k$ nearest neighbours are unlabelled; the gene is confidently not related to DR and is added to the set of reliable negatives. In case (b), the latter condition is not met; since the gene is not dissimilar enough to known DR-related genes, the gene is not added as a reliable negative, avoiding potential label noise during the training of the classifier in the second step of the PU Learning.

**Algorithm 1** Reliable Negatives selection by Nearest Neighbours

**Input:**
$D$: Set of training examples
$k$: Number of nearest neighbours
$t \in [0.5, 1]$: Threshold
$F$: Features to use
**Output:**
$RN$: Set of reliable negative training examples

**function** ReliableNegatives($P, U, k, t, F$)
  $P, U \leftarrow P, U$ considering only features F
  $D \leftarrow P \cup U$
  Initialize similarity matrix $S \in \mathbb{R}^{|D|\times|D|}$
  **for all** $(x_i, x_j)$; $x_i, x_j \in D$ **do** ▷ *Compute Jaccard distance for all pairs of examples*
    $S_{ij} \leftarrow \mathrm{J}(x_i, x_j)$ ▷ *Can cache S to avoid re-calculations of $S_{ij}$ across ReliableNegatives calls*
  $RN \leftarrow \emptyset$
  **for all** $x_i \in U$ **do**
    $T_k \leftarrow$ Top $k$ examples $x_j \in D \setminus x_i$ with highest similarity $S_{ij}$ to $x_i$
    $x_{max\_sim} \leftarrow$ The example $x_j \neq x_i$ with highest similarity $S_{ij}$ to $x_i$
    ▷ *If the % of unlabelled examples in the $Top_k$ nearest neighbours of $x_i$ exceeds a certain proportion and the single nearest neighbour of $x_i$ is unlabelled, add $x_i$ to the reliable negatives set* ◁
    **if** $\frac{|T_k \cap U|}{|T_k|} \geq t \subsetneq x_{max\_sim} \in U$ **then**
      $RN \leftarrow RN \cup x_i$
  **return** $RN$

pipeline. Following the state-of-the-art for this task [16], we perform a standard 10 × 5 nested-cross validation, where in each iteration of the outer loop, an inner loop is applied to the training set to search over the set of PU Learning hyperparameter combinations ($k$, $t$) detailed in Appendix A.

To increase the robustness of the method, we include an optional step in the training pipeline (see lines 14–16, 22–24 of Algorithm 2). Many of the feature sets of biological entities or genes are highly dimensional, and a similarity-based technique could make the PU Learning vulnerable to the *curse of dimensionality*, i.e. the similarity metric would not be informative due to the low signal-to-noise ratio arising from the high number of uninformative features [46]. To solve this for high-dimensional feature sets we add the option of, initially, training a classifier of choice, and then selecting the feature set $\mathcal{F}$ to be used in the nearest neighbours algorithm as the $n_f$ most important features for that classifier, according to a feature importance measure. Using a model-based filter feature selection for a subsequent algorithm is a well-studied and efficient method compared to more complex feature selection methodologies [47,48].

## 4. Experimental setup

This section covers the experimental setup used to evaluate the PU Learning methodology, compare it to existing approaches, and generate the ranking of promising candidate genes for DR-relatedness. We discuss the feature sets and classifiers used, details of the training and evaluation pipelines, and other implementation peculiarities.

### 4.1. Features and classifiers

In Section 2.1 we outlined that each ageing-related gene is represented as a vector of biological features, as well as a binary class label that indicating whether it has a known relation with DR (i.e. whether it belongs to $\mathcal{G}_{\mathrm{DR}\cap\mathrm{AGE}}$).

**Algorithm 2** Integration of PU Learning in Cross Validation

**Input:**
$D$: Set of examples
$k$, $t$: PU Learning hyperparameters
$n_f$: No. features to use in KNN

Split $D$ into 10 folds $D_1, ..., D_{10}$ ▷ *Outer CV*
**for** $i = 1, \dots, 10$ **do**
  $D_{Test} \leftarrow D_i$
  $D_{Train} \leftarrow D \setminus D_i$
  $AUC_{best} \leftarrow 0$
  **for all** hyperparameters combination $k, t$ **do**
    Split $D_{Train}$ into 5 folds $D_{Train_1}, ..., D_{Train_5}$
    **for** $i = 1, \dots, 5$ **do**
      $D_{Val} \leftarrow D_{Train_i}$
      $D_{Learn} \leftarrow D_{Train} \setminus D_{Train_i}$
      Train Model with $D_{Learn}$
      $F \leftarrow$ Top $n_f$ features in with highest Gini Importance in Model
      $P, U \leftarrow$ Positive and Unlabelled examples of $D_{Learn}$
      $RN \leftarrow$ **ReliableNegatives**($P$, $U$, $k$, $t$, $F$)
      Train Model with $P$ and $RN$ ▷ *Using all original features*
      Predict $D_{Val}$ with Model
    **if** $\text{F1}_{Model} > \text{F1}_{Best}$ **then** ▷ *Based on average AUC of 5 validation sets*
      $k_{best} \leftarrow k$, $t_{best} \leftarrow t$, $F1_{best} \leftarrow F1_{Model}$
  Train Model with $D_{Train}$ and hyperparams $H_{best}$
  $F \leftarrow$ Top $n_f$ features in with highest Gini Importance in Model
  $P, U \leftarrow$ Positive and Unlabelled examples of $D_{Train}$
  $RN \leftarrow$ **ReliableNegatives**($P$, $U$, $k_{best}$, $t_{best}$, $F$)
  Train Model with $P$ and $RN$ ▷ *Using all original features*
  Predict $D_{Test}$ using Model
  ▷ *Final performance is average F1, AUC and G.Mean of Model in the 10 test sets*

**Table 1**
Basic statistics of the two used datasets (feature types), with PathDIP [17] and GO [18] features. Among all ageing-related genes (i.e. $|\mathcal{G}_{\text{AGE}}|$ is the number of instances), the known DR-related genes ($\mathcal{G}_{\text{DR}\cap\text{AGE}} \subset \mathcal{G}_{\text{AGE}}$) constitute the subset of known positive instances or examples.

| Feature set | Features | Ageing-related genes ($\lvert\mathcal{G}_{\text{AGE}}\rvert$) | Known DR-related genes ($\lvert\mathcal{G}_{\text{DR}\cap\text{AGE}}\rvert$) | Feature sparsity (%) |
|---|---|---|---|---|
| PathDIP | 1640 | 986 | 110 | 98.39% |
| GO | 8640 | 1124 | 114 | 98.46% |

In this work, we focalize on the two feature sets originally found to be most relevant for the prediction of DR-relatedness by Magdaleno et al. [16]: PathDIP and GO features. Table 1 shows statistics of the two constructed datasets (one per feature type), namely the number of features, the number of genes (examples) in each class and the feature sparsity (percentage of 0-valued features) for each feature type. These two learning scenarios are created as follows:

- Using PathDIP [17], each gene is represented by binary features indicating whether or not it belongs to an specific PathDIP pathway; PathDIP, on its own, integrates information from multiple database sources such as Bio-Carta, REACTOME, UniProt, etc. For instance, if a gene has value 1 for the feature *KEGG.2*, then that gene has experimental evidence of relation to the *KEGG.2* pathway, which regulates *animal autophagy*.
- Using GO [18], each gene is represented by binary features indicating relation to a specific GO term or any of its descendants. For example, if a gene has value 1 for the feature *GO:0055114*, then that gene has experimental evidence of being related to the *Oxidation–reduction process*.

We followed Magdaleno et al.'s work to retrieve ageing-related genes, their features and their known DR-relatedness. To identify ageing-related human genes ($\mathcal{G}_{\text{AGE}}$), the GenAge database [49] was queried for genes that affected ageing phenotype or longevity if modulated in model organisms, and the obtained genes were mapped to their human orthologs using the OMA Orthology database [50] to obtain the final set of human ageing-related genes. The GenDR database [51] was queried to obtain DR-related genes in model organisms as those that affected the effectiveness of the DR-mediated ageing process in at least one wet lab experiment. Again, these DR-related genes were mapped to their human orthologs using the OMA Orthology database. The ageing-related genes were labelled as DR-related if they belonged to this set of DR-related genes, and PathDIP and GO features were retrieved for all ageing-related genes. For example, the ageing-related gene *MDH1* is labelled as positive (DR-related) because there is firm evidence it activates downstream targets of DR like SIR2 [52], while the ageing-related gene *PRKAB1* is unlabelled because there is an absence of evidence linking it to DR.

With regard to the classifiers used, our PU Learning method is classifier-agnostic, but we maintain the use of decision tree-based ensemble methods as done by Magdaleno et al. [16]; this is both to ensure fairness in the comparison with [16], and because tree-based ensembles are the State of the Art for tabular data, over options like Logistic Regression, Support Vector Machines or Neural Networks [24,53]. In particular, we employed CAT and BRF, the two classifiers with highest performance in the state-of-the-art approach for the problem [16]:

- CatBoost (CAT) [33] is a boosting-based ensemble classifier: each base learner is trained sequentially with instance weights determined by the errors of previous base learners in the sequence, progressively reducing the bias in the predictions [54].
- Balanced Random Forest (BRF) [30] is a bagging-based ensemble classifier: each base learner is trained independently using a bootstrap sample of the original data, and the high predictive accuracy is obtained through a reduction in variance of the errors, improving the unstable and inaccurate estimations of the weak base learners in isolation [54].

### 4.2. Evaluation details

In order to evaluate our proposed PU Learning method for the discovery of new candidate DR-related genes, we performed a dual evaluation, both on the surrogate binary classification task (predictive performance and computational cost) and on the proposed candidate ranking of the most promising novel DR-related genes.

To evaluate predictive performance we measure, for the existing (non-PU) approach [16] and our proposed PU Learning method, three relevant ML performance metrics: the F1 Measure of the positive class, the Geometric Mean (G. Mean), and the AUC-ROC [55], such that:

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \qquad G.Mean = \sqrt{Sensitivity * Specificity}$$
$$\text{AUC-ROC} = \text{Area Under the } Sensitivity \; vs. \; (1 - Specificity) \text{ Curve} \tag{9}$$

where Precision is the proportion of instances annotated as positives (DR-related) in the data among all instances predicted as positives, Recall (or Sensitivity) is the proportion of instances predicted as positives among the set of all instances annotated as positives in the data, and Specificity is the proportion of instances predicted as negatives (non-DR-related) among the set of all instances annotated as negatives in the data.

G. Mean and AUC-ROC are broadly used in PU Learning works for their global performance overview across both classes [56], and are used in the only existing work on DR-related genes identification [16]. However, AUC-ROC can be unreliable in PU tasks on datasets with class imbalance [57], but we keep it for fairness of comparison with [16]. We avoid the use of accuracy due to its inadequacy in scenarios with class imbalance [57], and we favour the use of F1 Score over its components Recall and Precision for this same reason.

We use the F1 Measure as the main performance indicator, as it is the most popular metric for PU Learning in the literature [56]. It is worth noticing that, because we utilized genuine PU data (i.e. we do not know the real labels of any of the unlabelled examples), all three metrics are an estimation of their real values: they evaluate the performance of the classifier predicting whether the example is a known positive ($Pr(g_i \in \mathcal{G}_{DR \cap AGE})$) rather than whether it is a real positive ($Pr(g_i \in \mathcal{G}_{DR \cap AGE^+})$). In this regard, the estimation of the F1 Measure of the positive class has desirable properties: Elkan and Noto [20] showed that (1) it will be a strict underestimation of the real F1 Measure, and (2) it will differ from the real F1 Measure by a constant factor, making it suitable for confidently comparing classifiers on discovery tasks.

With respect to the analysis of computational cost, we track the greenhouse gas emissions (grams of carbon dioxide equivalent or $gCO_2e$) of the DR-related gene identification pipelines of the existing (non-PU) method [16] and our PU-based Learning method, across the nested cross-validation training and inference pipeline, using *codecarbon* [58].

We also performe two qualitative evaluations of our best learned model: first, we analysed the most important features to predict DR-relatedness according to the classifier, based on the Mean Decrease in Impurity [59], and compared those features with the most important features identified in [16]. Second, we produce a ranking of the most promising ageing-related genes for novel DR-relatedness as predicted by our model. As formalized in Eq. (2), these most promising genes (i.e. those most likely to belong to $\mathcal{G}_{DR \cap AGE^+} \setminus \mathcal{G}_{DR \cap AGE}$) are defined by the model as those unlabelled examples (genes without known DR-relatedness) that are predicted as belonging to the positive class with the highest probability (see Eq. (4)). We use online resources (*Pubmed*, *Google Scholar*) to search for research linking a gene or its encoded protein to biological mechanisms potentially related to DR, for the most promising genes reported by our PU-based method and the existing approach by Magdaleno et al. [16].

### 4.3. Implementation details

This section covers technical details of the experiments performed to measure the performance of our PU Learning method and the existing (non-PU) Learning method from [16]:

- We implemented and performed all our experiments, using both the existing (non-PU) approach and our PU Learning approach, in a common Python framework. CatBoost and Balanced Random Forest are integrated through the *catboost*[1] and *imbalanced-learn*[2] packages, respectively. This framework[3] is publicly made available to the scientific community for transparency and further experimentation with these methods.
- For result reliability, the evaluation (nested cross-validation and ranking of candidate genes) of each method is averaged over 10 executions with random state seeds (14, 33, 39, 42, 727, 1312, 1337, 56 709, 177 013, 241 543 903).
- For the existing (non-PU) method, we replicated the original hyperparameter space and grid search procedure performed in [16]. For our PU Learning method, we tuned the PU Learning hyperparameters as described in Section 3. Appendix A details the values of tunable and non-tunable hyperparameters of all methods used in our experiments.
- We run all our experiments in a dedicated machine with 32 GB RAM, i9-10980 Intel©CPU, NVIDIA RTX 3070 GPU, and Windows 10 OS.

[1] https://catboost.ai/en/docs/.
[2] https://imbalanced-learn.org/stable/.
[3] https://github.com/Kominaru/DR_Gene_Prediction_XofN_PUL.

**Table 2**
Predictive performance of the original non-PU Learning method and our proposed PU Learning method in the identification of DR-related genes among ageing-related genes, considering two feature sets and two classifiers. Results represent average performance across 10 executions of the nested cross-validation procedure. For each metric, the best result is bolded and a dagger (†) represents statistical significance against all other results (in the other 7 rows) on two-tailed t-tests with $\alpha = 0.05$.

| Method | Features | Classifier | Performance (avg of 10 runs) | | |
|---|---|---|---|---|---|
| | | | AUC-ROC | G. Mean | F1 Score |
| Original | PathDIP | CAT | 0.829 | 0.717 | 0.522 |
| | | BRF | 0.825 | 0.752 | 0.450 |
| | GO | CAT | 0.832 | 0.654 | 0.463 |
| | | BRF | 0.827 | 0.755 | 0.377 |
| PU learning | PathDIP | CAT | 0.829 | 0.750 | **0.537**† |
| | | BRF | 0.815 | 0.728 | 0.381 |
| | GO | CAT | **0.838**† | 0.726 | 0.491 |
| | | BRF | 0.829 | **0.763**† | 0.380 |

## 5. Results

This Section covers the results of our PU Learning approach for detecting novel DR-related genes among ageing-related genes comparing its resuts with the results of the only existing approach for this task, a non-PU methodology recently introduced in [16]. We compare the two approaches' predictive performance results, computational cost, most important predictive features, and most promising candidate genes for novel DR-relatedness.

### 5.1. Results of computational experiments

Table 2 shows the predictive performance results (for the three metrics introduced in Section 4) of our PU Learning method and the existing, non-PU method [16]. For both methods, we consider two classifiers (CatBoost and Balanced Random Forest) and two biological feature sets (PathDIP pathways and GO terms).

The best results for each metric (AUC-ROC, G. Mean and F1 Score) are all obtained with our PU Learning method, showing statistical difference on two-tailed t-tests with $\alpha = 0.05$ against all other results. We highlight the relevance of the F1 Score, maximized by the PU Learning approach using {PathDIP, CAT}, as it is the most informative performance metric in PU Learning tasks [56]. As such, overall, our PU Learning method exhibits a stronger predictive performance than the original non-PU Learning approach by Magdaleno et al. [16].

Comparing the non-PU Learning and PU Learning approaches in each isolated {Features, Classifier} scenario, in both GO scenarios ({GO, CAT} and {GO, BRF}) the PU Learning approach outperformed the original non-PU method for all three performance metrics. In the {PathDIP, CAT} scenario, our PU-based model outperformed the non-PU Learning one in terms of F1 Score and G.Mean, exhibiting equivalent AUC-ROC. Only in one scenario ({PathDIP, BRF}) the usage of the PU Learning approach did not benefit the predictive performance; in this scenario, it seems that the benefits of a more reliable negative example labelling for training may not have compensated for the performance penalty that reducing the number of available training examples causes to a bagging-based method like BRF.

Overall, the best performance among all options in Table 2 was observed for the PU Learning approach in the {PathDIP, CatBoost} scenario, since it shows the statistically best result for F1, the most important metric -further analysed in Appendix B-, and exhibits competitive results in the other two metrics (AUC-ROC and G. Mean). As such, owing to an overall higher performance in these experiments involving two classifiers, two feature sets and three performance metrics, we can conclude that the proposed PU Learning approach is superior to state-of-the-art non-PU Learning methods in the identification of DR-related genes among ageing-related genes, and therefore should produce more reliable rankings of top candidates for DR-relatedness among genes without known relationship to DR.

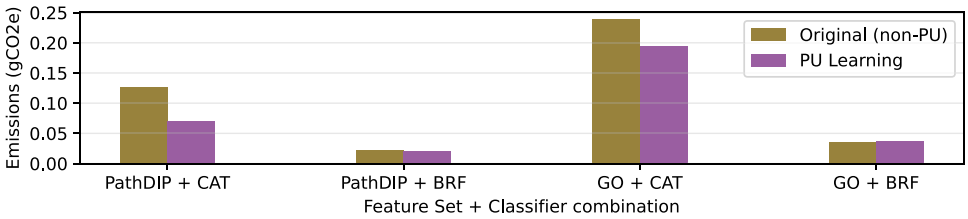


**Fig. 4.** Comparison of the computational cost (measured in grams of carbon dioxide equivalent ($gCO_2e$), lower is better) of Magdaleno et al.'s original non-PU approach and our proposed PU Learning-based approach for identification of new DR-related genes. Results are averaged over 10 complete executions of the nested cross-validation, involving training and inference procedures.

**Table 3**
The 5 most important features used to predict DR-relatedness using Magdaleno et al.'s non-PU Learning method [16] and our PU Learning algorithm, using PathDIP as feature set and CatBoost as classifier. The feature importance used is the Mean Decrease in Impurity based on the Gini Index. Each feature importance score was averaged across 10 executions of the nested 10 × 5 Cross Validation (CV) procedure, and it was normalized to the [0, 100] range. In each 10 × 5 CV, the feature importance is computed and averaged for each outer fold after the inner Cross Validation optimizes $k$ and $t$. Features common to both approaches are shown in italics.

| Original non-PU-Learning method | | | KNN-based PU-Learning method | | |
|---|---|---|---|---|---|
| Feature | Definition | Score | Feature | Definition | Score |
| *KEGG.2 (map04140)* | Autophagy - animal | 100.00 | *WikiPathways.37 (WP2884)* | NRF2 | 100.00 |
| *KEGG.30 (map04213)* | Longevity regulating pathway - multiple species | 45.35 | *KEGG.2 (map04140)* | Autophagy - animal | 72.52 |
| NetPath.23 (Pathway_BDNF) | Brain-derived neurotrophic factor | 38.95 | *KEGG.30 (map04213)* | Longevity regulating pathway - multiple species | 56.32 |
| REACTOME.10 (R-HSA-8953897) | Cellular responses to external stimuli | 37.79 | WikiPathways.57 (WP534) | Glycolysis and gluconeogenesis | 47.85 |
| *WikiPathways.37 (WP2884)* | NRF2 | 34.88 | EHMN.11 | Fructose and mannose metabolism | 43.97 |

With respect to the computational cost, Fig. 4 shows the average $CO_2$ emissions of the combined training and inference phases of the existing non-PU approach by Magdaleno et al. [16] and our proposed PU-based methodology. It can be observed that, despite the additional steps required for the PU Learning-based labelling of examples, this does not worsen the computational overhead of the identification pipeline. In fact, in the {PathDIP, CAT} scenario, where both the non-PU and our PU-based approach obtain highest predictive performance, our PU Learning approach can identify novel DR-related genes requiring ∼40% less greenhouse gas emissions. This is because, even if additional computations are required to select the reliable negative examples for the classifier, the resulting refined training set has fewer but better quality negative examples in it, prompting a more cost-effective learning of the model.

### 5.2. Analysis of most important predictive features

Table 3 shows the 5 most important features in the best models learnt for predicting DR-relatedness, i.e. the top-5 reported in [16] for the best model with the existing non-PU-Learning approach (left), and the top-5 in the best model learned with our PU Learning approach (right). This is a well-controlled comparison, as both sets were obtained in the {PathDIP, CatBoost} scenario. These feature importance values, as done in [16], were computed using the Mean Decrease in Impurity, which measures to what extent the splits with a given feature decrease the Gini Index across the trees in the ensemble [59].

In Table 3, three features ranked in the top-5 features for both approaches: KEGG pathway map04140 (Autophagy – animal), KEGG pathway map04213 (Longevity regulating pathway – multiple species), and WikiPathway WP2884 (NRF2). While "Longevity regulating pathway – multiple species" is a very broad KEGG pathway without clear relation to DR, there is good support in the literature for pathways map04140 and WP2884: autophagy inhibition tends to attenuate the anti-ageing effects of Calorie Restriction (CR), and a review on CR and autophagy has concluded that there is strong evidence that fasting and CR promote autophagy in a wide variety of tissues and organs [60]. Nuclear factor erythroid 2-related factor 2 (NRF2) is a transcription factor with activity regulated by DR that affects the expression of several enzymes with antioxidant and detoxifying functions [61].

Among the top-5 predictive features for the non-PU Learning approach reported in [16], NetPath Pathway_BDNF (Brain-derived neurotrophic factor) and Reactome R-HSA-8953897 (Cellular responses to external stimuli) are not among the top-5 features for our PU Learning method. Interestingly, it was noted in [16] that, among their reported top-5 features, only Pathway_BDNF and R-HSA-8953897 did not show significantly different degree of occurrence between genes with and without known DR relation, so the support for these features is weaker than for the other top-5 features. This was detected by the proposed PU Learning approach, which only ranked Pathway_BDNF and R-HSA-8953897 63rd and 34th, respectively, in terms of feature importance for predicting DR-relatedness.

Among the top-5 predictive features in our best PU Learning-based model, two are not among the top-5 features reported for the best model in [16]: WikiPathways WP534 (Glycolysis and gluconeogenesis) and EHMN.11 pathway (Fructose and mannose metabolism). Upon review of the existing scientific literature, there exists support for a significant role of these pathways in CR. Regarding WP534, in recent experiments quantifying the hepatic proteome of mice exposed to graded levels of CR (from 0% to 40%) for 3 months, one of the

**Table 4**
Top 7 candidate genes for novel DR-relatedness, i.e. genes without known DR association but with the highest likelihood of being DR related, for both the state-of-the-art non-PU Learning method and our proposed PU Learning algorithm. The DR-Probability is the output of the model for each gene, averaged across 10 executions of the 10 × 5 Cross Validation procedure. In each 10 × 5-CV, the DR-Probability is computed for the outer fold where the gene is in the test partition, after the inner Cross Validation optimizes $k$ and $t$. Genes common to both rankings are shown in italics.

| Original non-PU-Learning method | | KNN-based PU-Learning method | |
|---|---|---|---|
| Gene | DR-Probability | Gene | DR-Probability |
| GOT2 | 0.86 | *TSC1* | 0.97 |
| GOT1 | 0.85 | *GCLM* | 0.94 |
| *TSC1* | 0.85 | IRS1 | 0.93 |
| CTH | 0.85 | PRKAB1 | 0.92 |
| *GCLM* | 0.82 | PRKAB2 | 0.90 |
| *IRS2* | 0.80 | PRKAG1 | 0.90 |
| SENS2 | 0.80 | *IRS2* | 0.90 |

metabolic pathways most significantly stimulated by an increase in the level of CR was the glycolysis/gluconeogenesis pathway [62]. In addition, glycolysis and gluconeogenesis were considerably up-regulated in the kidney tissue of rats undergoing CR for 6 months [63]. Regarding EHMN.11, in a study to investigate the successful maintenance of weight loss in people after 8 weeks of low-calorie diet where people were classified as weight maintainers or weight regainers, an analysis of subcutaneous adipose tissue gene expression showed that the low-calorie diet caused a decrease in the fructose and mannose metabolism pathway in the weight regainer group [64].

When comparing the two sets of top-5 features identified by the two approaches in Table 3, it is worth recalling that the PU Learning-based model obtained better predictive performance than the non-PU model, as reported in Section 5.1. Therefore, it is reasonable to consider that the top-5 features identified by the PU Learning-based model are stronger, more reliable predictors of DR-relatedness than the top-5 features reported in [16].

### 5.3. Analysis of the most promising candidate DR-related genes

We employed the overall best method in Section 5.1 (the proposed PU learning method using PathDIP as feature set and CatBoost as classifier), to obtain the ranking of the most promising candidate ageing-related genes for novel DR-relatedness. For consistency with the evaluation scheme adopted in [16], which uses non-PU Learning methods, we report here the top-7 genes in the obtained ranking for each method.

Table 4 shows these top-7 most promising genes as identified in [16] (left) and by our PU-learning method (right); there is an overlap of 3 genes (*TSC1*, *GCLM*, *IRS2*) and 4 differing genes between the two approaches. Among the top-7 genes identified in [16], the 4 genes not occurring in the list of top-7 genes identified in this work are *GOT2*, *GOT1*, *CTH* and *SESN2*, ranked 10th, 11th, 23rd and 115th by our PU learning approach, respectively. Conversely, among the top-7 genes identified by the PU learning approach, 4 candidate DR-related genes do not occur in the list of top-7 genes identified in [16]: *IRS1*, *PRKAB1*, *PRKAB2* and *PRKAG1*. In the following paragraphs we discuss how evidence from the relevant literature supports the possible DR-relatedness of these 4 candidate DR-related genes identified by our PU Learning-based method (which should identify more reliable candidate genes owing to its higher predictive performance, as discussed in Section 5.1).

The PRKAB1 and PRKAB2 genes encode two isoforms of AMPK's regulatory $\beta$-subunit ($\beta$1 and $\beta$2) [65]; and PRKAG1 encodes an isoform of AMPK's regulatory subunit $\gamma$1 [66]. In previous research, fasting for 24 h increased the gene expression of AMPK $\beta$1- and $\beta$2-subunits in the hypothalamus of chicks [67]. In addition, experiments with a mouse model of Parkinson's disease showed that the hormone ghrelin mediated the neuroprotective effect of CR, and that the selective deletion of AMPK $\beta$1- and $\beta$2-subunits in dopamine neurons prevented ghrelin-induced AMPKD phosphorylation and neuroprotection [68].

PRKAG1 has shown to have an important role in the fasting-refeeding cycle associated with DR in killifish [69]: in young killifish, the fasting-refeeding cycle triggers an oscillatory regulation pattern in the expression of genes encoding the AMPK regulatory subunits $\gamma$1 and $\gamma$2, where fasting induces $\gamma$2 and suppresses $\gamma$1, whereas refeeding induces $\gamma$1 and suppresses $\gamma$2. This regulation pattern is blunted in old age, resulting in reduced PRKAG1 expression, which leads to chronic metabolic quiescence. Transgenic killifish with sustained AMPK-$\gamma$1 avoided that metabolic quiescence, leading to a more youthful feeding and fasting response in older killifish, with improved metabolic health. Hence, Ripa et al. [69] have proposed that the selective stimulation of AMPK-$\gamma$1 could be a good strategy to reinstate the beneficial response of a late-life DR through the maintenance of a correct refeeding response.

Regarding IRS1, experiments have shown a 109% in tyrosine-phosphorylated *IRS1* in insulin-treated muscles from rats on CR by comparison with control (fed *ad libitum*) rats [70], while experiments with mice lacking *IRS1* showed that the *IRS1*-encoded protein is not required for the CR-induced increase in insulin-stimulated glucose transport in skeletal muscle, and the absence of *IRS1* did not modify any of the measured characteristic adaptations of CR [71].

## 6. Conclusions

This work tackles the use of ML methods to identify new DR-related genes among ageing-related genes. The existing state-of-the-art method [16] treats as negative training examples (i.e. as non-DR-related genes) all genes without experimental evidence of DR-relatedness (unlabelled examples), introducing label noise to the training data and reducing the reliability of the identified candidate genes.

To address this limitation we propose a two-step, similarity-based PU Learning methodology for gene prioritization that creates a higher quality training set where all negative examples are reliable (confidently non-DR-related) to train the classifier. We compared our PU-based method against the state-of-the-art, non-PU-based methodology [16] for the task of identifying DR-related genes among ageing-related genes.

We show that our PU Learning approach significantly ($p < 0.05$) outperforms the state-of-the-art methodology [16] for our target task, in terms of F1 Score, G.Mean and AUC-ROC. Moreover, our method lowers the computational cost of the gene identification task by up to ~40% in the best-performing scenario, as it generates a set of negative training examples that is smaller but has higher quality (it is more reliable), allowing a more cost-effective learning.

We use our best model (trained in the {PathDIP, CatBoost} scenario) to generate a ranking of candidate DR-related genes; and compare it to the ranking reported in [16] obtained without PU Learning. We identify several new potentially DR-related genes (e.g. IRS1, PRKAB1, PRKAB2, PRKAG1). Our new list of candidates is, from a ML standpoint, more reliable than the one curated by Magdaleno et al. since it was generated by a model with significantly higher predictive performance. Moreover, a curation of the scientific literature of these genes supports the potential relation of the top promising genes with the mechanisms of DR.

Regarding future research, we identify different avenues: (1) the validation of the DR-related role of the identified genes in wet-lab experiments, (2) the combination of multiple and other biological feature sets with our PU-based method to improve the quality of predictions, and (3) exploration of other types of classifiers to further evaluate our PU Learning method in gene prioritization scenarios.

**Table A.1**
Full relation of classifier (non-tunable) and PU Learning (tunable) hyperparameters used in our experiments. For CAT and BRF, all hyperparameter values are identical to those in [16], and hyperparameters not stated in this table use the default values in *python* libraries *catboost v1.2.2* and *imbalanced-learn v0.12.0*.

| Hyperparameter group | Parameter | Value(s) |
|---|---|---|
| PU Learning (tunable) | No. of neighbours | $k \in 3, 5, 8$ |
| | Selection threshold | $k = 3 : t \in {}^{2}/_{3}, 1$<br>$k = 5 : t \in {}^{4}/_{5}, 1$<br>$k = 8 : t \in {}^{6}/_{8}, {}^{7}/_{8}, 1$ |
| BRF | Estimators | $n = 500$ |
| | Sampling strategy | $\omega_{us} = 1$ |
| | Replacement | $r = True$ |
| CAT | Estimators | $n = 500$ |

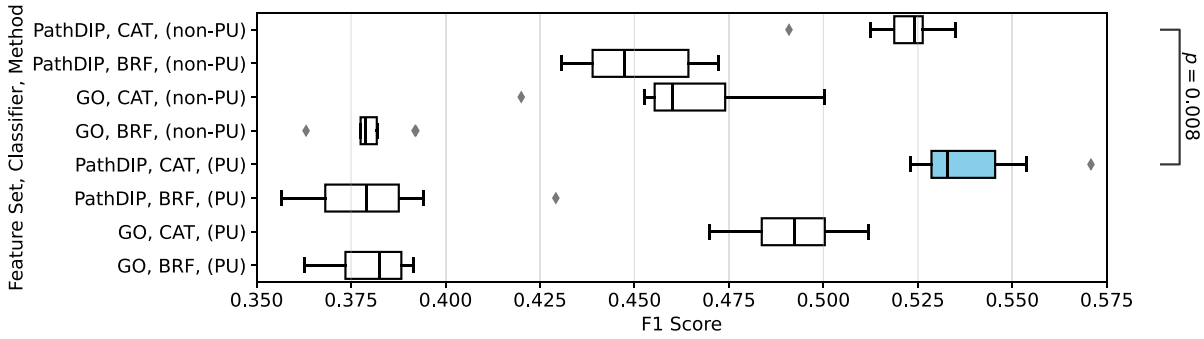


**Fig. B.1.** Details of the F1 Score across 10 executions of the nested cross-validation for the existing method [16] and our PU Learning-based proposal. For the best method, highlighted in blue (PU Learning on the {PathDIP, CAT} scenario), the $p$-value of the paired t-test against the best-performing scenario of the non-PU method is shown.

## CRediT authorship contribution statement

**Jorge Paz-Ruza:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Alex A. Freitas:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Amparo Alonso-Betanzos:** Writing – review & editing, Supervision, Conceptualization. **Bertha Guijarro-Berdiñas:** Writing – review & editing, Supervision, Conceptualization.

## 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.

## Acknowledgements

This research work has been funded by MICIU/AEI/10.13039/501100011033 and ESF+ (grant FPU21/05783), MICIU/AEI/10.13039/501100011033/ and *ERDF A way of making Europe* (grant PID2019-109238GB-C22), and by the Xunta de Galicia (Grant ED431C 2022/44) with the European Union ERDF funds. CITIC, as Research Center accredited by Galician University System, is funded by "Consellería de Cultura, Educación e Universidade from Xunta de Galicia", supported in an 80% through ERDF Operational Programme Galicia 2021–2027, and the remaining 20% by "Secretaría Xeral de Universidades" (Grant ED431G 2023/01).

## Appendix A. Detailed experimental hyperparameters

See Table A.1.

## Appendix B. Detailed F1 score results

See Fig. B.1.

## References

[1] C. López-Otín, L. Galluzzi, J.M.P. Freije, F. Madeo, G. Kroemer, Metabolic control of longevity, Cell 166 (4) (2016) 802–821.
[2] L. Partridge, The new biology of ageing, Phil. Trans. R. Soc. B 365 (1537) (2010) 147–154.
[3] L. Guarente, C. Kenyon, Genetic pathways that regulate ageing in model organisms, Nature 408 (6809) (2000) 255–262.
[4] D. Melzer, L.C. Pilling, L. Ferrucci, The genetics of human ageing, Nature Rev. Genet. 21 (2) (2020) 88–101.
[5] J. Most, V. Tosti, L.M. Redman, L. Fontana, Calorie restriction in humans: an update, Ageing Res. Rev. 39 (2017) 36–45.
[6] M. Das, I. Gabriely, N. Barzilai, Caloric restriction, body fat and ageing in experimental models, Obes. Rev. 5 (1) (2004) 13–19.
[7] E. Kirk, D.N. Reeds, B.N. Finck, M.S. Mayurranjan, B.W. Patterson, S. Klein, Dietary fat and carbohydrates differentially alter insulin sensitivity during caloric restriction, Gastroenterology 136 (5) (2009) 1552–1560.
[8] J.G. Wood, B. Rogina, S. Lavu, K. Howitz, S.L. Helfand, M. Tatar, D. Sinclair, Sirtuin activators mimic caloric restriction and delay ageing in metazoans, Nature 430 (7000) (2004) 686–689.
[9] E. Bergamini, G. Cavallini, A. Donati, Z. Gori, The role of autophagy in aging: Its essential part in the anti-aging mechanism of caloric restriction, Ann. New York Acad. Sci. 1114 (1) (2007) 69–78.
[10] G. López-Lluch, P. Navas, Calorie restriction as an intervention in ageing, J. Physiol. 594 (8) (2016) 2043–2060.
[11] T.S. de Carvalho, Calorie restriction or dietary restriction: how far they can protect the brain against neurodegenerative diseases? Neural Regen. Res. 17 (8) (2022) 1640–1644.
[12] C. Angermueller, T. Pärnamaa, L. Parts, O. Stegle, Deep learning for computational biology, Mol. Syst. Biol. 12 (7) (2016) 878.
[13] K.A. Shastry, H. Sanjay, Machine learning for bioinformatics, in: Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications, Springer, 2020, pp. 25–39.
[14] C. Kern, P. Manescu, M. Cuffaru, C. Au, A. Zhange, H. Wang, A.F. Gilliat, M. Ezcurra, D. Gems, Machine learning predicts lifespan and underlying causes of death in aging C. elegans, 2024, bioRxiv.
[15] F. Fabris, J.P.d. Magalhães, A.A. Freitas, A review of supervised machine learning applied to ageing research, Biogerontology 18 (2017) 171–188.
[16] G.D. Vega Magdaleno, V. Bespalov, Y. Zheng, A.A. Freitas, J.P. De Magalhaes, Machine learning-based predictions of dietary restriction associations across ageing-related genes, BMC Bioinform. 23 (2022) 1–28.
[17] S. Rahmati, M. Abovsky, C. Pastrello, I. Jurisica, pathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis, Nucleic Acids Res. 45 (D1) (2017) D419–D426.

[18] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature Genet. 25 (1) (2000) 25–29.

[19] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.

[20] C. Elkan, K. Noto, Learning classifiers from only positive and unlabeled data, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 213–220.

[21] F. Li, S. Dong, A. Leier, M. Han, X. Guo, J. Xu, X. Wang, S. Pan, C. Jia, Y. Zhang, et al., Positive-unlabeled learning in bioinformatics and computational biology: a brief review, Brief. Bioinform. 23 (1) (2022) bbab461.

[22] R. Kiryo, G. Niu, M.C. Du Plessis, M. Sugiyama, Positive-unlabeled learning with non-negative risk estimator, Adv. Neural Inf. Process. Syst. 30 (2017).

[23] J. Bekker, J. Davis, Learning from positive and unlabeled data: A survey, Mach. Learn. 109 (4) (2020) 719–760.

[24] L. Grinsztajn, E. Oyallon, G. Varoquaux, Why do tree-based models still outperform deep learning on typical tabular data? Adv. Neural Inf. Process. Syst. 35 (2022) 507–520.

[25] F. Fabris, A.A. Freitas, New KEGG pathway-based interpretable features for classifying ageing-related mouse proteins, Bioinformatics 32 (19) (2016) 2988–2995.

[26] R. Oughtred, C. Stark, B.-J. Breitkreutz, J. Rust, L. Boucher, C. Chang, N. Kolas, L. O'Donnell, G. Leung, R. McAdam, et al., The BioGRID interaction database: 2019 update, Nucleic Acids Res. 47 (D1) (2019) D529–D541.

[27] G. Consortium, The GTEx consortium atlas of genetic regulatory effects across human tissues, Science 369 (6509) (2020) 1318–1330.

[28] S. van Dam, T. Craig, J.P. de Magalhães, GeneFriends: a human RNA-seq-based gene and transcript co-expression database, Nucleic Acids Res. 43 (D1) (2015) D1124–D1132.

[29] J. Rainer, EnsDb. Hsapiens. v86, Bioconductor, 2017.

[30] C. Chen, A. Liaw, L. Breiman, Using Random Forest to Learn Imbalanced Data, Technical Report 666, Department of Statistics, UC Berkley, 2004, URL: http://xtf.lib.berkeley.edu/reports/SDTRWebData/accessPages/666.html.

[31] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[32] X.-Y. Liu, J. Wu, Z.-H. Zhou, Exploratory undersampling for class-imbalance learning, IEEE Trans. Syst. Man Cybern. B 39 (2) (2008) 539–550.

[33] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, Adv. Neural Inf. Process. Syst. 31 (2018).

[34] Y. Zheng, H. Peng, X. Zhang, Z. Zhao, X. Gao, J. Li, DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions, BMC Bioinform. 20 (2019) 1–12.

[35] W. Lan, J. Wang, M. Li, J. Liu, Y. Li, F.-X. Wu, Y. Pan, Predicting drug–target interaction using positive-unlabeled learning, Neurocomputing 206 (2016) 50–57.

[36] C. Kılıç, M. Tan, Positive unlabeled learning for deriving protein interaction networks, Netw. Model. Anal. Health Inform. Bioinform. 1 (2012) 87–102.

[37] H. Song, B.J. Bremer, E.C. Hinds, G. Raskutti, P.A. Romero, Inferring protein sequence-function relationships with large-scale positive-unlabeled learning, Cell Syst. 12 (1) (2021) 92–101.

[38] D. Zhang, H. Xing, D. Liu, M. Han, P. Cai, H. Lin, Y. Tian, Y. Guo, B. Sun, Y. Le, et al., Discovery of toxin-degrading enzymes with positive unlabeled deep learning, ACS Catal. 14 (2024) 3336–3348.

[39] K. He, Y. Wang, X. Xie, D. Shao, A multi-task positive-unlabeled learning framework to predict secreted proteins in human body fluids, Complex Intell. Syst. 10 (1) (2024) 1319–1331.

[40] B. Liu, Y. Dai, X. Li, W.S. Lee, P.S. Yu, Building text classifiers using positive and unlabeled examples, in: Third IEEE International Conference on Data Mining, IEEE, 2003, pp. 179–186.

[41] S. Sellamanickam, P. Garg, S.K. Selvaraj, A pairwise ranking based approach to learning with positive and unlabeled examples, in: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, 2011, pp. 663–672.

[42] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inform. Theory 13 (1) (1967) 21–27.

[43] P. Teisseyre, K. Furmańczyk, J. Mielniczuk, Verifying the selected completely at random assumption in positive-unlabeled learning, 2024, arXiv preprint arXiv: 2404.00145.

[44] P. Jaccard, Étude comparative de la distribution florale dans une portion des Alpes et des Jura, Bull. Soc. Vaudoise Sci. Nat. 37 (1901) 547–579.

[45] T.T. Tanimoto, Elementary mathematical theory of classification and prediction, 1958.

[46] N. Altman, M. Krzywinski, The curse (s) of dimensionality, Nature Methods 15 (6) (2018) 399–400.

[47] K. Grabczewski, N. Jankowski, Feature selection with decision tree criterion, in: Fifth International Conference on Hybrid Intelligent Systems, HIS'05, IEEE, 2005, pp. 6–pp.

[48] C.a. Ratanamahatana, D. Gunopulos, Feature selection for the naive bayesian classifier using decision trees, Appl. Artif. Intell. 17 (5–6) (2003) 475–487.

[49] J.P. de Magalhaes, O. Toussaint, GenAge: a genomic and proteomic network map of human ageing, FEBS Lett. 571 (1–3) (2004) 243–247.

[50] A.M. Altenhoff, C.-M. Train, K.J. Gilbert, I. Mediratta, T. Mendes de Farias, D. Moi, Y. Nevers, H.-S. Radoykova, V. Rossier, A. Warwick Vesztrocy, et al., OMA orthology in 2021: website overhaul, conserved isoforms, ancestral gene order and more, Nucleic Acids Res. 49 (D1) (2021) D373–D379.

[51] R. Tacutu, D. Thornton, E. Johnson, A. Budovsky, D. Barardo, T. Craig, E. Diana, G. Lehmann, D. Toren, J. Wang, et al., Human ageing genomic resources: new and updated databases, Nucleic Acids Res. 46 (D1) (2018) D1083–D1090.

[52] E. Easlon, F. Tsang, C. Skinner, C. Wang, S.-J. Lin, The malate–aspartate NADH shuttle components are novel metabolic longevity regulators required for calorie restriction-mediated life span extension in yeast, Genes Dev. 22 (7) (2008) 931–944.

[53] R. Shwartz-Ziv, A. Armon, Tabular data: Deep learning is not all you need, Inf. Fusion 81 (2022) 84–90.

[54] P. Bühlmann, Bagging, boosting and ensemble methods, in: Handbook of Computational Statistics: Concepts and Methods, Springer, 2012, pp. 985–1022.

[55] N. Japkowicz, M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011.

[56] J.D. Saunders, A.A. Freitas, Evaluating the predictive performance of positive-unlabelled classifiers: a brief critical review and practical recommendations for improvement, ACM SIGKDD Explor. Newsl. 24 (2) (2022) 5–11.

[57] M. Bekkar, H.K. Djemaa, T.A. Alitouche, Evaluation measures for models assessment over imbalanced data sets, J. Inf. Eng. Appl. 3 (10) (2013).

[58] B. Courty, V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, Jérémy Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, M. Stęchły, C. Bauer, Lucas-Otavio, JPW, MinervaBooks, mlco2/codecarbon: v2.4.1, 2024, http://dx.doi.org/10.5281/zenodo.11171501.

[59] G. Louppe, L. Wehenkel, A. Sutera, P. Geurts, Understanding variable importances in forests of randomized trees, Adv. Neural Inf. Process. Syst. 26 (2013).

[60] M. Bagherniya, A. Butler, G. Barreto, A. Sahebkar, The effect of fasting or calorie restriction on autophagy induction: A review of the literature, Ageing Res. Rev. 47 (2018) 183–197.

[61] A. Vasconcelos, N. Dos Santos, C. Scavone, C. Munhoz, Nrf2/ARE pathway modulation by dietary energy regulation in neurological disorders, Front. Pharmacol. 10 (2019).

[62] L. Wang, D. Derous, X. Huang, S. Mitchell, A. Douglas, D. Lusseau, Y. Wang, J. Speakman, Impact of graded calorie restriction on protein expression in the liver, J. Gerontol. A 78 (7) (2023) 1125–1134.

[63] J. Chen, C. Velalar, R. Ruan, Identifying the changes in gene profiles regulating the amelioration of age-related oxidative damages in kidney tissue of rats by the intervention of adult-onset calorie restriction, Rejuvenation Res. 11 (4) (2008) 757–763.

[64] D. Mutch, H. Tune, M. Pers, R. Temanni, P. Marquez-Quinones, C. Holst, J. Martinez, et al., A distinct adipose tissue gene expression response to caloric restriction predicts 6-mo weight maintenance in obese subjects, Am. J. Clin. Nutr. 94 (6) (2011) 1399–1409.

[65] O. Katwan, F. Alghamdi, T. Almabrouk, S. Mancini, S. Kennedy, J. Oakhill, J. Scott, I. Salt, AMP-activated protein kinase complexes containing the Beta 2 regulatory subunit are up-regulated during and contribute to adipogenesis, Biochem. J 476 (2019) 1725–1740.

[66] H. An, Y. Wang, C. Qin, M. Li, A. Maheshwari, L. He, The importance of the AMPK gamma 1 subunit in metformin suppression of liver glucose production, Sci. Rep. 10 (2020) 10482.

[67] L. Lei, Z. Lixian, Effect of 24h fasting on gene expression of AMPK, appetite regulation peptides and lipometabolism related factors in the hypothalamus of broiler chicks, Asian-Aust. J. Anim. Sci. 25 (9) (2012) 1300–1308.

[68] J. Bayliss, M. Lemus, R. Stark, V. Santos, A. Thompson, D. Rees, S. Galic, J. Elsworth, B. Kemp, J. Davies, Z. Andrews, Ghrelin-AMPK signaling mediates the neuroprotective effects of calorie restriction in Parkinson's disease, J. Neurosci. 36 (10) (2016) 3049–3063.

[69] R. Ripa, E. Ballhysa, J. Steiner, R. Laboy, A. Annibal, N. Hochhard, C. Latza, L. Dolfi, C. Calabrese, A. Meyer, M. Polidori, R. Muller, A. Antebi, Refeeding-associated AMPK$\gamma$1 complex activity is a hallmark of health and longevity, Nat. Aging 3 (2023) 1544–1560.

[70] D. Dean, G. Cartee, Calorie restriction increases insulin-stimulated tyrosine phosphorylation of insulin receptor and insulin receptor substrate-1 in rat skeletal muscle, Acta Psyshiol. Scand. 169 (2000) 133–139.

[71] A. Gazdag, C. Dumke, C. Kahn, G. Cartee, Calorie restriction increases insulin-stimulated glucose transport in skeletal muscle from IRS-1 knockout mice, Diabetes 48 (1999) 1930–1936.