Interpretable Ensembles of Classifiers for Uncertain Data with Bioinformatics Applications

Marcelo Rodrigues de Holanda Maia, Alexandre Plastino, Alex A. Freitas, and João Pedro de Magalhães

Abstract—Data uncertainty remains a challenging issue in many applications, but few classification algorithms can effectively cope with it. An ensemble approach for uncertain categorical features has recently been proposed, achieving promising results. It consists in biasing the sampling of features for each model in an ensemble so that less uncertain features are more likely to be sampled. Here we extend this idea of biased sampling and propose two new approaches: one for selecting training instances for each model in an ensemble and another for sampling features to be considered when splitting a node in a Random Forest training. We applied these approaches to classify ageing-related genes and predict drugs' side effects based on uncertain features representing protein-protein and protein-chemical interactions. We show that ensembles based on our proposed approaches achieve better predictive performance. In particular, our proposed approaches improved the performance of a Random Forest based on the most sophisticated approach for handling uncertain data in ensembles of this kind. Furthermore, we propose two new approaches for interpreting an ensemble of Naive Bayes classifiers and analyse their results on our datasets of ageing-related genes and drug's side effects.

Index Terms—Classification, interpretability, data uncertainty, bioinformatics, the biology of ageing.

1 Introduction

Data uncertainty can be categorised into existential uncertainty, which occurs when the existence of some data record is uncertain, and value uncertainty, which can be further categorised into class-label uncertainty or feature-value uncertainty. This work addresses feature-value uncertainty, which occurs when some feature values in a data record (instance) are not precisely known. This uncertainty can naturally arise due to the limited precision of data collection technology, particularly in bioinformatics or biomedical domains. An uncertain feature value is usually represented by a probability distribution on the corresponding feature's domain.

It has been shown that incorporating information on uncertainty into classification algorithms can improve predictive performance [1], [2], [3], [4], [5], but this is still an underexplored research topic, particularly for categorical features, since most previous methods focus on uncertain numerical features. Hence, this work proposes new ensemble methods for coping with uncertain categorical features.

We focus on ensemble methods that learn many base classifiers independently on random subsets of the original training set and then aggregate the base classifiers' predictions. In general, such ensemble methods usually have better predictive performance and are more robust to slight data variations than any single base classifier. In particular, Bagging methods randomly sample subsets of the instances in the dataset with replacement (which is called bootstrap sampling) [6], and Random Subspaces methods randomly sample subsets of the features in the dataset [7].

An ensemble approach for uncertain categorical features, named Biased Random Subspaces (BRS), has recently been proposed by Maia et al. [3]. It consists in biasing the random sampling of features for each model in an ensemble, based on the principle that features with lower uncertainty degrees should have better class-discrimination potential since the confidence about their actual values is higher.

Relying on the same hypothesis, this work extends the idea of biased random sampling by proposing two new approaches for building ensembles of classifiers that cope with uncertain categorical features. The first is a Biased Bootstrap (BB) approach for selecting training instances for each model in an ensemble. The second is a Biased Splitting (BS) approach for sampling features to be considered when splitting a node while building the trees of a Random Forest.

We evaluate our proposed approaches by using them to build Naive Bayes (NB) classifiers ensembles and Random Forests and performing experiments on ten classification datasets with real uncertain information. This uncertainty consists of feature values' probability distributions extracted from real-world databases – unlike previous work, which typically used datasets with artificially generated uncertainty [1], [2], [4], [5].

Out of these 10 datasets, 4 were also used in [3]. In these datasets, each instance is an ageing-related gene. The class labels indicate whether a gene has a pro-longevity or anti-longevity effect on a particular organism's lifespan, as recorded in the GenAge database [8], and the features represent protein-protein interaction (PPI) information. The feature uncertainty is represented by probabilities of interactions between two proteins, obtained from the STRING database [9].

The other 6 datasets are introduced in this work. In these

- *M.R.H. Maia is with Instituto de Computação, Universidade Federal Fluminense, Niterói, Brazil and Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brazil.*
  E-mail: mmaia@ic.uff.br; marcelo.h.maia@ibge.gov.br
- *A. Plastino is with Instituto de Computação, Universidade Federal Fluminense, Niterói, Brazil*
- *A.A. Freitas is with the School of Computing, University of Kent, Canterbury, UK*
- *J.P. de Magalhães is with the Genomics of Ageing and Rejuvenation Lab, University of Birmingham, Birmingham, UK*

datasets, each instance is a drug (chemical). The class labels indicate whether or not a drug has a particular side effect, as recorded in the SIDER database [10], and the features represent protein-chemical interaction (PCI) information. The feature uncertainty is represented by probabilities of interactions between a chemical (drug) and a protein, obtained from the STITCH database [11].

We report two types of results. First, we compare the predictive performance of the ensemble methods using the aforementioned uncertainty-handling approaches (BRS, BB and BS) against the accuracy of the corresponding baseline ensemble methods (without the BRS, BB and BS approaches), using two well-known predictive performance measures: the Area Under the ROC curve (AUROC) and the geometric mean of Sensitivity and Specificity [12]. In general, the proposed 'biased-sampling' ensemble methods outperformed the baseline 'unbiased-sampling' methods.

Additionally, we report the results of interpreting the best ensemble models learned from the ageing-related datasets. Model interpretation is an increasingly important topic in machine learning [13], and it has led to novel biological insights in bioinformatics domains [14], [15].

Although a single NB classifier is naturally interpretable, interpreting an NB ensemble is not trivial. We are aware of only one approach in the literature for interpreting an NB ensemble: transforming it into a single NB model by linear approximation (with some loss of predictive accuracy) and then interpreting that NB model [16], [17].

Hence, we propose two new approaches for interpreting an NB ensemble as a further contribution. The first measures the importance of a feature based on conditional probability differences, whereas the second is a more sophisticated approach based on finding a minimal set of features that is sufficient to preserve the class predicted for an instance (so that changes to the values of other features in that instance do not change the class predicted by the model). We use these two interpretation approaches for NB ensembles and a conventional feature importance measure for random forests to learn feature rankings for the ageing-related datasets, identifying the most important features for predicting such genes' effects on an organism's lifespan.

In summary, this paper extends the initial work from [3] by providing four new contributions. First, we propose two new approaches – Biased Bootstrap (BB) and Biased Splitting (BS) – for learning from uncertain categorical features, which complement the Biased Random Subspace (BRS) one introduced in [3]. Second, in [3] the BRS approach was used to create NB ensembles only; whilst in this paper, we create several types of ensembles, including NB ensembles with the BB approach and random forests with the BS and BB approaches. Third, the experiments in [3] used only 4 datasets of ageing-related genes, whilst in this paper, we use 10 datasets: those 4 datasets and 6 new datasets for predicting drugs' side effects. Fourthly, this paper introduces two approaches for interpreting an ensemble of Naive Bayes classifiers, whilst no such interpretation was attempted in [3]. We also use these interpretation approaches to analyse the best models learned from the ageing-related datasets, discussing the results from the perspective of the biology of ageing as an interdisciplinary contribution.

2 METHODS

2.1 Definitions

Let $F = \{f_1, f_2, \ldots, f_m\}$ be the set of predictive features, where $m \geq 1$, and $C = \{c_1, c_2, \ldots, c_q\}$ be the set of classes, where $q \geq 2$. The domain of a feature f_j is $dom(f_j)$. A dataset $D = \{(X_1, y_1), (X_2, y_2), \ldots, (X_n, y_n)\}$ consists of n labelled instances. Each instance in D, identified by an index i, is associated with a feature vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ and a class label $y_i \in C$. In the classification problem, the objective is to construct a model from D capable of predicting the class of an unlabelled instance given its corresponding feature vector.

Let $U \subseteq F$ be the set of uncertain features, assumed to be categorical in this work. The domain of a categorical feature f_j is a finite set of values $dom(f_j) = \{v_{j1}, v_{j2}, \ldots, v_{j|dom(f_j)|}\}$, $|dom(f_j)| \geq 2$. If f_j is not uncertain, its value x_{ij} for an instance i is represented by a single value. Otherwise, it is a discrete probability distribution represented by a probability vector P_{ij}, i.e.:

$$x_{ij} = \begin{cases} x_{ij} \in dom(f_j), & \text{if } f_j \in F \setminus U \\ P_{ij} = (p_{ij1}, p_{ij2}, \ldots, p_{ij|dom(f_j)|}), & \text{otherwise} \end{cases}$$

where, if $f_j \in U$, $p_{ijk} \in [0, 1]$ represents the probability that x_{ij} takes the value v_{jk} and $\sum_{k=1}^{|dom(f_j)|} p_{ijk} = 1$.

2.2 Coping with uncertainty in ensemble models

Recently, an ensemble approach for coping with uncertainty in categorical features, named Biased Random Subspaces (BRS), has been proposed by Maia et al. [3]. It consists in biasing the random sampling of features for each model in an ensemble. Here we extend the idea of biased random sampling to two new approaches: a Biased Bootstrap (BB) approach for selecting training instances for each model in an ensemble (which can be used with any bagging-based ensemble algorithm) and a Biased Splitting (BS) approach for sampling features to be considered when splitting a node while building the trees of a Random Forest.

From [3], we have the definition of the bias value for a feature f_j, given by:

$$b_{*j} = \left(1 - \frac{1}{|I \setminus M_{*j}|} \sum_{i \in I \setminus M_{*j}} E_{ij}\right) \times \frac{|I \setminus M_{*j}|}{|I|}$$

where $I = \{1, 2, \ldots, n\}$ is the set of indices of all instances in D, M_{*j} is the set of indices of instances in D with a missing value for the feature f_j, and E_{ij} is the normalized entropy of the probability distribution represented by P_{ij} if f_{ij} is an uncertain feature (or zero, otherwise), that is:

$$E_{ij} = \begin{cases} \frac{\sum_{k=1}^{|dom(f_j)|} p_{ijk} log(p_{ijk})}{log(1/|dom(f_j)|)}, & \text{if } f_j \in U \\ 0, & \text{otherwise} \end{cases}$$

The feature bias values are normalized over all features, defining a probability distribution $B = (\beta_1, \beta_2, \ldots, \beta_m)$, where a probability β_j associated with a feature f_j is given by $\beta_j = b_{*j} / (\sum_{l=1}^{m} b_{*l})$.

Instead of the default uniform distribution from the general Random Subspaces strategy, the BRS approach uses

the probability distribution B to sample the features to train each base classifier in the ensemble.

Random Forests usually do not sample features before generating each tree. Nonetheless, they sample a subset of features to be considered candidate features when splitting each tree node. Hence, we propose the Biased Splitting (BS) approach for this sampling, which uses the probability distribution B to sample the candidate features.

Finally, we propose the Biased Bootstrap (BB) approach for instance sampling, which is analogous to the BRS and BS approaches for feature sampling. Hence, we define the bias value for an instance identified by index i, given by:

$$b_{i*} = \left(1 - \frac{1}{|F \setminus M_{i*}|} \sum_{f_j \in F \setminus M_{i*}} E_{ij}\right) \times \frac{|F \setminus M_{i*}|}{|F|}$$

where M_{i*} is the set of features in D with a missing value for instance i.

The instance bias values are normalized over all instances, defining a distribution $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$, where a probability γ_i associated with an instance identified by index i is given by $\gamma_i = b_{i*}/(\sum_{l=1}^{n} b_{l*})$.

The BB approach uses probability distribution Γ to sample the instances for training each base classifier.

2.3 Interpreting NB ensembles via conditional probabilities

The first approach we propose for interpreting an ensemble of NB classifiers relies on the influence that a feature value $x_{ij} \in dom(f_j)$ has for determining the most likely class to be predicted for an instance by a single NB classifier, which we compute as an importance score. We then combine the importance scores from all the classifiers into the ensemble's importance scores.

This approach does not address feature uncertainty. It assumes the base models of the ensemble are standard NB classifiers. Therefore, in this context, x_{ij} is always represented by a single value.

Given a feature vector $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ associated with an unlabelled instance identified by index i, an NB classifier predicts the class $y \in C$ that maximizes the value given by $P(y|X_i) \propto P(y) \prod_{j=1}^{m} P(x_{ij}|y)$.

We first present our definition of importance for binary classifiers, where $C = \{c_1, c_2\}$. The importance of a feature value x_{ij} in a given NB classifier is estimated by the following difference between conditional probabilities:

$$Diff(x_{ij}, c_1, c_2, e) = |P(x_{ij}|c_1) - P(x_{ij}|c_2)|$$

where e is the classifier for which the difference is computed. Clearly, the higher the difference in the class-conditioned probability of a feature value between the two class labels, the more importance (influence) that feature value will have for determining the most likely class to be assigned to the testing instance.

For datasets with more than two class labels, this idea can be generalised by summing the differences between all pairs of class labels in C:

$$Importance(x_{ij}, C, e) = \sum_{r=1}^{q-1} \sum_{s=r+1}^{q} Diff(x_{ij}, c_r, c_s, e)$$

The importance of a feature value x_{ij} for an ensemble of NB classifiers is computed by averaging its importance across all classifiers in the ensemble. However, different NB classifiers will generally use different feature subsets (due to the random subspaces approach). Intuitively, other things being equal, the larger the number of classifiers in the ensemble that use a feature, the larger the importance of the value of that feature. Therefore, we assume that, if a classifier e_u does not use a feature f_j, then $Importance(x_{ij}, C, e_u) = 0$. Hence, we define the ensemble-wide importance as:

$$Importance(x_{ij}, C) = \frac{\sum_{u=1}^{t} Importance(x_{ij}, C, e_u)}{t}$$

where e_u is the u-th classifier and t is the total number of classifiers in the ensemble.

This equation is appropriate when the predicted class returned by the ensemble is computed by a simple majority vote of all classifiers, i.e., all classifiers have the same weight in the voting. If weighted voting is used instead (where the weight of a vote is proportional to the classifier's confidence in its prediction), then the importance equation could be easily modified to compute a correspondingly weighted average over the t classifiers.

Finally, once the importance value has been computed for all feature values x_{ij}, we rank all feature values in decreasing order of importance. Then a user (domain expert) can focus on interpreting the top-ranked feature values, i.e., the most important ones for predicting the class variable in the ensemble.

Note that for binary domain features, where $dom(f_j) = \{v_{j1}, v_{j2}\}$, the importance computed for both values will be the same due to the complementarity of the probabilities in use. Given two class labels c_r and c_s such that $c_r \in C$, $c_s \in C$ and $c_r \neq c_s$, the following relations apply:

$$P(v_{j1}|c_r) = 1 - P(v_{j2}|c_r) \tag{1}$$

$$P(v_{j1}|c_s) = 1 - P(v_{j2}|c_s) \tag{2}$$

The difference of the class-conditioned probabilities of the value v_{j1} between c_r and c_s for a classifier e would be:

$$Diff(v_{j1}, c_r, c_s, e) = |P(v_{j1}|c_r) - P(v_{j1}|c_s)| \tag{3}$$

By replacing (1) and (2) in (3), we obtain:

$$\begin{aligned} Diff(v_{j1}, c_r, c_s, e) &= |(1 - P(v_{j2}|c_r)) - (1 - P(v_{j2}|c_s))| \\ &= |1 - P(v_{j2}|c_r) - 1 + P(v_{j2}|c_s)| \\ &= |P(v_{j2}|c_s) - P(v_{j2}|c_r)| \\ &= |P(v_{j2}|c_r) - P(v_{j2}|c_s)| \\ &= Diff(v_{j2}, c_r, c_s, e) \end{aligned}$$

Therefore, for binary domain features (like the PPI and PCI features in our datasets), the importance computed for both values in the domain is the same. Then, the importance computed for any of the values in a feature's domain can be interpreted as that feature's importance.

Algorithm 1. MinimalSufficientSet(X_i, e)

$c_i \leftarrow$ class predicted by e for X_i
$SuppSet \leftarrow \{f_j | P(x_{ij}|c_i) > P(x_{ij}|y), \forall y \in C \setminus \{c_i\}\}$
$S \leftarrow SuppSet$
Calculate $Importance(x_{ij}, C, e)$ for all $f_j \in S$
$SortedFeats \leftarrow$ SortByImportance($SuppSet$)
$e' \leftarrow e$
for each feature f_j in $SortedFeats$ **do**
  Remove f_j from e'
  $c'_i \leftarrow$ class predicted by e' for X_i
  if $c'_i = c_i$ **then** $S \leftarrow S \setminus \{f_j\}$
  else Exit loop
return S

2.4 Interpreting NB ensembles via minimal sufficient features

The approach based on conditional probability differences measures the importance of each feature value separately, ignoring its importance in the context of all other feature values. This is consistent with NB assuming that each feature is independent of all others conditioned on the class variable, but it does not directly measure the influence of a feature value on class prediction. Naive Bayes makes class predictions using the formula: $P(y|X_i) \propto P(y)\prod_{j=1}^{m} P(x_{ij}|y)$. Therefore, whether or not a conditional probability will make a difference in the choice of the predicted class depends on the entire set of conditional probabilities and the prior class probability.

Hence, we propose a second approach that considers sets of feature values. The principle we rely on is the same adopted in the definitions of *anchors* by Ribeiro et al. [18] and *minimal sufficient factors* by Watson et al. [19]. We seek to find, for each instance, a minimal set of features sufficient to preserve the class prediction, such that changes to the other feature values of the instance would not change the class predicted by the model.

Note that the larger the value of $Importance(x_{ij}, C, e)$, the higher the influence of x_{ij} for a class prediction in model e, but even the feature with the highest $Importance$ value may still not be sufficient for a given prediction.

However, we can use this notion to find a minimal sufficient set of features. The basic idea is to sort all features in increasing order of their $Importance$ values and then identify the minimal set of top features in that sorted list which, together, are sufficient for preserving the class prediction made by the classifier.

Let c_i be the class predicted by the classifier for instance i. A feature value x_{ij} is said to "support" the prediction of c_i if and only if $P(x_{ij}|c_i) > P(x_{ij}|y), \forall y \in C \setminus \{c_i\}$. That is, the feature value x_{ij} becomes more likely if instance i has class c_i than if that instance has another class. Naturally, when searching for a minimal sufficient set of features, we only need to consider feature values that support the class predicted by the classifier.

A method for identifying a minimal sufficient set of features for the class prediction for a given instance is presented in Algorithm 1.

Based on the sufficiency criterion, we define a measure of the importance of a feature f_j for a classifier e given the set of instances X, denoted $SImportance(f_j, e, X)$, as the proportion of instances in X for which f_j is in the minimal sufficient set returned by Algorithm 1.

The importance of a feature for the entire ensemble is computed by simply averaging its importance across all classifiers in the ensemble:

$$SImportance(f_j, X) = \frac{\sum_{u=1}^{t} SImportance(f_j, e_u, X)}{t}$$

where e_u is the u-th classifier and t is the total number of classifiers in the ensemble.

2.5 Datasets

We have evaluated the proposed approaches on real data from two application domains. The first domain is the classification of ageing-related genes regarding their effect on the lifespan of an organism, which may be positive (pro-longevity) or negative (anti-longevity). From this domain, we used the 4 datasets generated by Maia et al. [3], which integrate data from the GenAge database (Build 20) [8] and the STRING database (Version 11.0) [9]. Each dataset contains data regarding ageing-related genes of one of the 4 major model organisms from the GenAge database: *C. elegans* (roundworm), *D. melanogaster* (fruit fly), *M. musculus* (mouse), and *S. cerevisiae* (baker's yeast). Each feature in these datasets refers to a protein-protein interaction (PPI) extracted from the STRING database.

The second domain involves the prediction of drugs' side effects. For this domain, we have generated 6 new datasets by integrating data from the SIDER database (Version 4.1) [10] and the STITCH database (Version 5.0) [11]. The SIDER database contains information on marketed medicines and their recorded side effects (adverse drug reactions). STITCH is a database of protein-chemical interactions (PCI) that stem from computational predictions, knowledge transfer between organisms, and interactions aggregated from other databases.

Each side-effect dataset refers to one of the 6 most frequent side effects in the SIDER database: nausea, headache, dermatitis, rash, vomiting and dizziness. Each instance in these datasets refers to a drug and consists of uncertain features referring to PCIs and a binary class variable indicating whether the corresponding drug has the side effect represented in the dataset (positive) or not (negative). Each PCI feature refers to one protein and has a binary domain, indicating whether or not an interaction between the corresponding chemical (drug) and the protein referred by the feature has been observed. As uncertain features, they are represented by probability distributions.

A value x_{ij} of an uncertain binary feature f_j for an instance i in a dataset is represented by a probability distribution $P_{ij} = (p_{ij1}, p_{ij2})$, where p_{ij1} and p_{ij2} are the complementary probabilities of x_{ij} taking each of the two values in $dom(f_j)$. Therefore, each probability distribution representing a PPI or PCI feature value is encoded by a single value p_{ij}, and $P_{ij} = (p_{ij}, 1 - p_{ij})$. In our datasets, p_{ij} is the confidence score (interaction probability) obtained from the STRING or STITCH databases for the corresponding PPI or PCI features, respectively.

Table 1 presents detailed information about the datasets. They are particularly challenging for having many features,

TABLE 1
Information about the Datasets

Dataset	Instances	Features	Missing Values (%)	Class (%)	
				Neg.	Pos.
AG-Worm	763	9692	93.8	66.3	33.7
AG-Fly	185	3883	88.4	37.3	62.7
AG-Mouse	82	4216	78.4	37.8	62.2
AG-Yeast	382	4274	90.3	88.0	12.0
SE-Nausea	1394	9096	97.5	15.4	84.6
SE-Headache	1394	9096	97.5	21.7	78.3
SE-Dermatitis	1394	9096	97.5	23.2	76.8
SE-Rash	1394	9096	97.5	23.8	76.2
SE-Vomiting	1394	9096	97.5	24.0	76.0
SE-Dizziness	1394	9096	97.5	27.3	72.7

a small number of instances, and a very high percentage of missing values (when there is no information regarding a specific interaction in the STRING or STITCH databases). To avoid overfitting, we have discarded PPI and PCI features with low support (annotating less than 10 instances). As usual in the literature using PPIs as features for classifying genes, we represent missing values as zeros.

2.6 Ensemble methods

We consider three baseline ('unbiased-sampling') ensemble methods: two kinds of NB ensembles and one Random Forest (RF). Although these baseline methods do not use an uncertainty-based random sampling bias, they handle uncertainty at the individual classifiers' level.

The baseline NB ensembles, ENB-NV and ENB-EV, were also used by Maia et al. [3]. In ENB-NV (Ensemble of NB classifiers with Numeric Values), the NB classifiers treat each uncertain value (an interaction probability) as a numeric value and assume that the feature values' probability distributions are Gaussian. In ENB-EV (Ensemble of NB classifiers with Expected Values), the NB classifiers binarise each uncertain value into an expected value using the threshold 0.5 and consider multivariate Bernoulli distributions for the data.

To the best of our knowledge, there is no work in the literature coping with uncertain categorical features using RFs. Among papers coping with uncertain numerical features using RF (or single decision trees), the most sophisticated approach is distributing fractions of examples (DFE) over the child nodes when splitting a node on an uncertain feature [4], [20], [21], [22]. The DFE approach provided RFs with mildly positive results in [21] and mildly negative results in [22]. Although only evaluated in the literature for numerical features, RFs using the DFE approach (RF-DFE) also apply to categorical ones. Since RF-DFE is the most sophisticated available RF that can handle uncertain categorical features, we use it as our baseline RF.

The baseline ensembles perform a standard unbiased sampling of instances and features. For each of them, we build three versions by incorporating different combinations of our proposed approaches: BB, BRS and BB+BRS for NB ensembles; BB, BS and BB+BS for Random Forests.

We have coded the algorithms[1] by extending available implementations from the scikit-learn library [23]. We have set the number of base classifiers (NB or decision trees) in each ensemble to 500 and the number of instances used to build each one to n. The number of features sampled (to train each NB classifier or to split a tree node in an RF) has been set to $\sqrt{m}$. Since our RFs distribute fractions of examples using instance weights, instead of defining a minimum number of instances required to be at a leaf node, we defined the minimum sum of instance weights required at a leaf node as 0.01 (1% of the number of instances).

1. The source-code and datasets used in this work are available at https://github.com/marcelorhmaia/interpretable-ensembles-for-ucd

2.7 Predictive performance measures

We have assessed the predictive performance of the algorithms using two metrics: the Area Under the Receiver Operating Characteristic curve (AUROC) and the geometric mean of sensitivity and specificity (G-mean) [12].

We evaluated each algorithm using the well-known 10-fold cross-validation. Furthermore, we have assessed the statistical significance of the differences in the predictive performance measures between each pair of algorithms, using a paired Wilcoxon signed-rank test for each dataset, with a significance level of 0.05.

2.8 Simpson's paradox

Let X be a binary feature, taking values x_1 or x_2, and Y be the class variable in a dataset. Let the dataset's instances be divided into two groups: those with $X = x_1$ and those with $X = x_2$. Consider a class label of interest, y_1 (e.g., the pro-longevity or anti-longevity label in our ageing-related datasets). Let $P(y_1|x_1)$ and $P(y_1|x_2)$ denote the conditional probabilities of y_1 for the corresponding groups of instances. Consider the scenario where each group of instances is further divided according to the values of another binary feature Z, called a confounder, taking values z_1 or z_2 (in our datasets, the confounders are binary, but this condition could be relaxed). Simpson's paradox occurs when $P(y_1|x_1) > P(y_1|x_2)$ and $P(y_1|x_1, z_j) < P(y_1|x_2, z_j)$, for $j \in \{1, 2\}$ or vice-versa: $P(y_1|x_1) < P(y_1|x_2)$ and $P(y_1|x_1, z_j) > P(y_1|x_2, z_j)$, for $j \in \{1, 2\}$. That is, the paradox occurs if the conditional probability of the class label of interest y_1 'increases' ('decreases') from the group where $X = x_1$ to the group where $X = x_2$ but, surprisingly, the conditional probability of y_1 'decreases' ('increases') from the former to the latter group, both for instances with $Z = z_1$ and instances with $Z = z_2$ [24], [25], [26]. Hence, the paradox shows a reversal of the direction of association between the values of a feature and the probability of a class label of interest in the context of a confounder.

3 Results

3.1 Assessing predictive performance

3.1.1 Experiment 1

This experiment evaluated NB ensembles using NB-NV as base classifiers (i.e., Gaussian NB with Numeric Values of features). We have compared four ensembles: the baseline ensemble of NB-NVs, denoted ENB-NV; and three biased-sampling ensembles, combining ENB-NV with the biased sampling approaches, denoted ENB-NV+BB, ENB-NV+BRS and ENB-NV+BB+BRS.

Table 2 shows the AUROC and G-mean results on the 10 datasets. The best values (for each metric and dataset) are in bold. The last row shows each ensemble's average rank (per metric). Superscript symbols indicate the statistically significant advantages (SSA). Based on these results, ENB-NV+BRS is the best ensemble regarding AUROC and G-mean, with an average rank of 1.6 for both metrics.

3.1.2 Experiment 2

This experiment evaluated NB ensembles using NB-EV (which binarises uncertain values into Expected Values) as base classifiers. Again, we have compared four ensembles: the unbiased-sampling baseline ensemble of NB-EVs, denoted ENB-NV; and three biased-sampling ensembles combining ENB-EV with the biased sampling approaches, denoted ENB-EV+BB, ENB-EV+BRS and ENB-EV+BB+BRS. Table 3 presents the results for this group of ensembles. ENB-EV+BRS is the best ensemble for AUROC and G-mean, with the average ranks of 1.4 and 1.6, respectively.

3.1.3 Comparing the best NB ensembles

Table 4 presents the results for ENB-NV+BRS and ENB-EV+BRS, the best ensembles from Experiments 1 and 2, respectively. ENB-EV+BRS obtained the best AUROC results, with an average rank of 1.4, whereas ENB-NV+BRS was the best for G-mean, with an average rank of 1.0.

3.1.4 Experiment 3

This experiment evaluated four RFs: RF-DFE (distributing fractions of examples among child nodes), the most sophisticated available RF that can handle uncertain categorical features; and three biased-sampling RFs, combining RF-DFE with the biased sampling approaches: RF-DFE+BB, RF-DFE+BS, RF-DFE+BB+BS. Table 5 presents the results. RF-DFE+BB obtained the best AUROC results, with an average rank of 2.0; whereas RF-DFE+BS obtained the best G-mean results, also with an average rank of 2.0.

3.1.5 Comparing the best NB ensembles and the best RFs

This last comparison aimed at determining the best overall method regarding each of the AUROC and G-mean metrics. Table 6 presents the results. RF-DFE+BB obtained the best AUROC results (average rank: 1.1), whereas RF-DFE+BS obtained the best G-mean results (average rank: 1.3).

In general, the results of these three experiments support the hypothesis that the BB, BRS and BS approaches can improve the predictive performance of ensembles on uncertain data. A limitation of the BB approach was that it produced poor results when applied to NB ensembles. However, the BB and BS approaches proposed in this paper worked well with RFs, and they obtained the best overall results since RF-DFE+BB and RF-DFE+BS were the best overall ensembles for the AUROC and G-mean metrics, respectively, outperforming RF-DFE.

3.2 Identifying the top-ranked PPI features

We have applied our two proposed approaches for interpreting NB ensembles to the four ageing-related datasets to identify the top-ranked PPI features for classification.

Among all NB ensembles evaluated in our experiments, ENB-EV+BRS and ENB-NV+BRS have achieved the best overall AUROC and G-mean values, respectively. We have selected ENB-EV+BRS for model interpretation since it uses binarised features, facilitating interpretation.

We have trained a model applying ENB-EV+BRS to each dataset's whole set of instances. Then, we used this model to produce rankings of features in decreasing order of importance.

Table 7 presents the top-10 features for each organism (dataset) regarding the importance measure based on conditional probabilities. For each feature, columns 2–5 present its importance-based rank, the corresponding protein ID from the STRING database, and the corresponding gene's symbol and name. Columns 6 and 7 present the absolute and relative frequencies of value 1 (considering the threshold of 0.5 used) for the corresponding feature in the instances of the Pro-longevity and Anti-longevity classes. The last column shows whether or not the feature is involved in occurrences of Simpson's paradox [24], [25] (discussed in Subsection 3.4).

Interestingly, out of the 40 top-ranked PPI features in Table 7, 15 represent ribosomal proteins, namely 5 of the top-10 PPI features for the worm dataset and all the top-10 PPI features for the yeast dataset. It is also worth observing in Table 7 the relative frequency of genes (dataset instances) of each class (Pro- vs Anti-longevity) that interact with the gene associated with each of these 15 ribosomal proteins (PPI features). In all those 15 table rows, the relative frequency of Anti-longevity genes interacting with the corresponding ribosomal protein is substantially higher than that of Pro-longevity genes interacting with that ribosomal protein. The relative frequency differences are clearly striking for 9 of the 10 yeast PPI features in the table, which have a frequency of 0% for Pro-longevity genes and frequencies varying from 14.0% to 18.2% for Anti-longevity genes.

Despite this strong pattern, none of these 9 yeast ribosomal proteins is included in GenAge [8] – the most comprehensive database of ageing-related genes. By itself, this strong pattern does not allow us to conclude that those 9 ribosomal proteins have an anti-longevity effect on yeast, which in principle could be confirmed only via appropriate biological experiments. However, the pattern seems strong enough to justify further investigation of some of those 9 ribosomal proteins in future work.

Among the 5 ribosomal proteins in Table 7 for worm, 4 (rps-0, rps-5, rps-11, rpl-3) are included in the GenAge database. Actually, among the top-10 PPI features, 8 are associated with genes included in GenAge – the exceptions are atp-1 and rps-30.

Regarding the top-10 PPI features for mice, only the top-ranked one, igf1, is included in GenAge. Interestingly, igf1 is annotated as having an "unclear" effect on longevity in GenAge, whilst its closely related igf1r (igf1 receptor) is annotated as Anti-longevity. In Table 7, the relative frequency of Anti-longevity genes interacting with the igf1 gene is 51.6%, which is much larger than the relative frequency of such interaction in the Pro-longevity class: 19.6%.

Finally, out of the top-10 PPI features for fly in Table 7, 4 are associated with genes included in GenAge, namely Sod1, FOXO, Sod2, park. Among the 6 genes not included in GenAge, there are 3 heat shock proteins. Two of them,

TABLE 2
Experiment 1 Results

Dataset	AUROC (%)				G-mean (%)			
	ENB-NV	ENB-NV +BB	ENB-NV +BRS	ENB-NV +BB+BRS	ENB-NV	ENB-NV +BB	ENB-NV +BRS	ENB-NV +BB+BRS
AG-Worm	71.46	72.32	**72.33**	72.26	57.94	57.34	**60.83***†	60.55†
AG-Fly	65.62	65.94	65.03	**66.37**	57.31	57.61	**59.79**	58.37
AG-Mouse	66.73†	63.66	**68.51**	66.96	57.09	59.94	56.28	**63.13**
AG-Yeast	61.62	**63.81***	61.22	62.74‡	57.76	**60.51**§	58.90§	54.66
SE-Nausea	58.89	56.53	**65.16***†§	51.10	20.47	24.97	**28.97***	25.60
SE-Headache	56.35§	54.64§	**56.92**§	51.75	**53.61**†§	23.20	43.22†§	24.45
SE-Dermatitis	58.06§	56.05§	**59.83**†§	51.86	20.81	36.11	**54.16***†§	21.48
SE-Rash	56.98§	55.24§	**59.78**†§	51.10	21.12	40.63§	**54.94***†§	20.40
SE-Vomiting	60.64†§	58.24§	**65.19***†§	53.69	23.18	**34.78**	33.55*§	28.70
SE-Dizziness	61.63†§	55.21	**65.01***†§	51.63	24.73	49.86	**57.00***†§	23.21
Avg. Rank	2.5	2.7	**1.6**	3.2	3.2	2.4	**1.6**	2.8

*SSA (vs ENB-NV), †SSA (vs ENB-NV+BB), ‡SSA (vs ENB-NV+BRS), §SSA (vs ENB-NV+BB+BRS)

TABLE 3
Experiment 2 Results

Dataset	AUROC (%)				G-mean (%)			
	ENB-EV	ENB-EV +BB	ENB-EV +BRS	ENB-EV +BB+BRS	ENB-EV	ENB-EV +BB	ENB-EV +BRS	ENB-EV +BB+BRS
AG-Worm	**76.91**†‡§	75.39§	74.81§	73.49	36.27	38.14	49.27*†	**54.68***†‡
AG-Fly	64.05	66.24	**69.00**	67.91	28.28	26.09	28.06	**30.10**
AG-Mouse	69.26	**70.03**	69.30	68.92	41.59	35.58	**46.82**	35.10
AG-Yeast	75.55	75.77	76.79	**78.30**	22.23	32.62	47.78*†	**67.77**†‡
SE-Nausea	50.45	46.81	**57.06**†	46.79	21.84	18.03	**25.68***†	19.20
SE-Headache	52.25†§	48.86	**54.86***†§	48.26	24.64	18.57	**25.76**†§	20.17
SE-Dermatitis	54.55†§	51.07	**58.57***†§	52.00	22.12	17.89	**23.51**†	21.64†
SE-Rash	53.65†§	49.19	**57.71***†§	49.76	21.72	17.59	**23.06**†	21.24†
SE-Vomiting	56.43†§	46.27	**65.91***†§	47.95†	**15.99**	14.21	15.00	14.18
SE-Dizziness	57.72†§	46.07	**65.20***†§	46.90	**23.78**†§	14.23	23.72†	21.58†
Avg. Rank	2.4	3.1	**1.4**	3.1	2.2	3.6	**1.6**	2.6

*SSA (vs ENB-EV), †SSA (vs ENB-EV+BB), ‡SSA (vs ENB-EV+BRS), §SSA (vs ENB-EV+BB+BRS)

TABLE 4
Comparison of the Best NB Ensembles

Dataset	AUROC (%)		G-mean (%)	
	ENB-NV +BRS	ENB-EV +BRS	ENB-NV +BRS	ENB-EV +BRS
AG-Worm	72.33	**74.81**	**60.83***	49.27
AG-Fly	65.03	**69.00**	**59.79***	28.06
AG-Mouse	68.51	**69.30**	**56.28***	46.82
AG-Yeast	61.22	**76.79***	**58.90**	47.78
SE-Nausea	**65.16**	57.06	**28.97**	25.68
SE-Headache	**56.92**	54.86	**43.22***	25.76
SE-Dermatitis	**59.83**	58.57	**54.16***	23.51
SE-Rash	**59.78***	57.71	**54.94***	23.06
SE-Vomiting	65.19	**65.91**	**33.55***	15.00
SE-Dizziness	65.01	**65.20**	**57.00***	23.72
Avg. Rank	1.6	**1.4**	**1.0**	2.0

*SSA

Hsp70Ab and Hsp70Aa, have a relative frequency of only 5.8% for the Anti-longevity class, with a much larger relative frequency of 19.0% and 19.8%, respectively, for the Pro-longevity class.

Table 8 presents the top-10 features for each organism regarding the importance measure based on minimal sufficient features. Most features from Table 7 are also in Table 8: 8 features for worm, 9 for fly, 6 for mouse and 6 for yeast.

As RF-DFE+BB+BS achieved the best predictive performance on the ageing-related datasets, we also produced feature rankings using it. We have used the default RF feature importance measure from the scikit-learn's implementation, based on the Gini index[2]. Table 9 presents the top-10 features for each organism. Consistently with Tables 7 and 8, many of the top-10 PPI features for worm and yeast in Table 9 represent ribosomal proteins, which are mainly associated with the anti-longevity class.

3.3 Confirming the relevance of the top-ranked features for the biology of ageing

Overall, the top-ranked features fit well with current knowledge on longevity/ageing and previous similar analyses, as follows. The top-ranked features for worms include daf-16, a key regulator of longevity in worms. Mutations in daf-16 suppress the longevity effects caused by several

2. The Gini feature importance measure is biased towards features with many values [27]. However, since all features in our datasets are binary, this limitation is not an issue in this work.

TABLE 5
Experiment 3 Results

Dataset	AUROC (%)				G-mean (%)			
	RF-DFE	RF-DFE +BB	RF-DFE +BS	RF-DFE +BB+BS	RF-DFE	RF-DFE +BB	RF-DFE +BS	RF-DFE +BB+BS
AG-Worm	**76.62**†‡§	75.41	74.86	74.73	52.95	56.55*	55.83*	**60.22***†‡
AG-Fly	67.87	**71.81**	67.27	69.57	**66.87**†	57.43	64.42†	65.91
AG-Mouse	68.38	67.68	70.66	**72.60**	**60.65**	56.63	59.14	56.63
AG-Yeast	77.44	78.75	78.51	**79.11**	51.99	51.03	52.46	**57.30**
SE-Nausea	69.33	**69.99**	69.08	69.52	55.91†§	16.40	**57.26**†§	16.40
SE-Headache	**68.55**§	66.66	67.59	66.88	**59.39**†§	20.18	57.96†§	22.62
SE-Dermatitis	65.11‡	66.44*‡	62.89	**66.86***‡	52.20†§	20.90	**54.75***†§	19.39
SE-Rash	64.83‡	**66.26**‡	63.45	66.11*‡	53.17†§	21.35	**55.05***†§	19.66
SE-Vomiting	68.02	**68.32**	68.15	67.83	51.05	58.69	56.20*	**59.29**
SE-Dizziness	68.55	69.02	67.74	**69.67**†‡	50.98	54.66	56.81*	**60.23**†
Avg. Rank	2.7	**2.0**	3.2	2.1	2.4	3.1	**2.0**	2.3

*SSA (vs RF-DFE), †SSA (vs RF-DFE+BB), ‡SSA (vs RF-DFE+BS), §SSA (vs RF-DFE+BB+BS)

TABLE 6
Comparison of the Best NB Ensembles and the Best RFs

Dataset	AUROC (%)		G-mean (%)	
	ENB-EV +BRS	RF-DFE +BB	ENB-NV +BRS	RF-DFE +BS
AG-Worm	74.81	**75.41**	**60.83**	55.83
AG-Fly	69.00	**71.81**	59.79	**64.42***
AG-Mouse	**69.30**	67.68	56.28	**59.14**
AG-Yeast	76.79	**78.75**	**58.90**	52.46
SE-Nausea	57.06	**69.99***	28.97	**57.26***
SE-Headache	54.86	**66.66***	43.22	**57.96***
SE-Dermatitis	58.57	**66.44***	54.16	**54.75***
SE-Rash	57.71	**66.26***	54.94	**55.05***
SE-Vomiting	65.91	**68.32**	33.55	**56.20***
SE-Dizziness	65.20	**69.02***	**57.00***	56.81
Avg. Rank	1.9	**1.1**	1.7	**1.3**

*SSA

mutations [28]. Other top genes in worms include various ribosomal proteins, which is not surprising given that they control translation, which has been strongly associated with longevity regulation in worms and other organisms [29]. Mitochondrial genes are also among the top hits, which also fits current knowledge of the role of mitochondria in ageing and longevity regulation [30]. Lastly, one of the top genes is atg-7, an autophagy regulator that is a major longevity pathway in invertebrates, including in worms [31].

There are heat shock proteins and genes related to stress response among the top genes in flies. This fits well with the long-established observation that stress resistance is important for healthy ageing and longevity [32]. There are also antioxidant enzymes, like superoxide dismutase, thioredoxin and glutathione peroxidase; in invertebrates and flies, in particular, antioxidant protection has long been considered important for longevity [33]. Interestingly, in flies, we see repair pathways and mechanisms that protect against stress among top features, particularly with a high frequency of pro-longevity interactions. In yeast, most top features are ribosomal proteins, which, as mentioned earlier, have been related to longevity regulation in model organisms.

In mice, the top gene is igf-1, with a strong anti-longevity frequency. The growth hormone/insulin/IGF1 pathway is the major longevity pathway in mammals [34], [35], so this result fits our knowledge of longevity well. Also, other players in the pathway like forkhead box proteins, Pik3cd, Ins2 and Irs2 are among the top predictions. Some brain and neuronal factors (e.g., Src) are also among the top features, which could fit GH/IGF1's neuroendocrine regulation [34]. Alternatively, they could be related to ageing changes in the brain. As Src is not in GenAge, it could be an interesting target for future studies.

Overall, the top-ranked features fit nicely into pro- and anti-longevity pathways enriched in GenAge [31]. Of note, in worms, ribosomal proteins and mitochondrial proteins involved in oxidative phosphorylation have been previously found enriched in anti-longevity processes [31], while in flies, responses to oxidative stress (like antioxidant enzymes) are among enriched pro-longevity processes. In mice, the insulin signalling pathway is a top enriched anti-longevity pathway [31], in line with our results.

3.4 Detecting occurrences of Simpson's paradox

When interpreting an association between each top-ranked feature and a class label (pro- or anti-longevity), it is important to consider that the direction of that association might be misleading due to Simpson's paradox. By direction of association we mean whether the presence of a PPI is associated with an increased probability of the pro-longevity or, conversely, the anti-longevity class.

Simpson's paradox occurs when the direction of an association between two variables X and Y at the population (aggregated) level is reversed in all the sub-groups produced by partitioning that population according to the values of a third variable, Z, called a confounder [24], [25]. In other words, the direction of the association between variables X and Y is reversed when conditioning on each value of the confounder Z. In our classification task, X and Z are predictive features, whilst Y is the class variable.

Table 10 shows an example of this paradox in the fly dataset. Looking only at the aggregated data in the first row of the table, ignoring the interaction between the values of the Hsp83 and Hsc70-4 PPI features, we would conclude

TABLE 7
Top-10 PPI Features in the ENB-EV+BRS Model According to the Importance Measure Based on Conditional Probabilities

Organism	Feat. Rank	STRING ID	Gene Symbol	Protein Name	Freq. in Pro-long.	Freq. in Anti-long.	Parad.
Worm	1	R13H8.1h	daf-16	Forkhead box protein O	40 (15.6%)	43 (8.5%)	no
	2	F42G8.12	isp-1	Cytochrome b-c1 complex subunit Rieske, mitochondrial	9 (3.5%)	58 (11.5%)	no
	3	C34E10.6.1	atp-2	ATP synthase subunit beta, mitochondrial	6 (2.3%)	58 (11.5%)	no
	4	T05E11.1	rps-5	40S ribosomal protein S5	5 (1.9%)	60 (11.9%)	no
	5	F40F11.1.2	rps-11	Ribosomal protein, small subunit	4 (1.6%)	53 (10.5%)	no
	6	C26F1.4.2	rps-30	40S ribosomal protein S30	5 (1.9%)	43 (8.5%)	no
	7	B0250.1	rpl-2	60S ribosomal protein L8	3 (1.2%)	44 (8.7%)	no
	8	B0393.1.1	rps-0	40S ribosomal protein SA	6 (2.3%)	51 (10.1%)	no
	9	H28O16.1a	H28O16.1	ATP synthase subunit alpha, mitochondrial	6 (2.3%)	56 (11.1%)	no
	10	Y56A3A.19	Y56A3A.19	Acyl carrier protein	2 (0.8%)	51 (10.1%)	no
Fly	1	FBpp0082516	Hsc70-4	Heat shock 70 kDa protein cognate 4	34 (29.3%)	9 (13.0%)	no
	2	FBpp0305736	Sod	Superoxide dismutase [Cu-Zn]	30 (25.9%)	4 (5.8%)	no
	3	FBpp0081956	Hsp70Ab	Heat shock protein 70Ab	22 (19.0%)	4 (5.8%)	no
	4	FBpp0293589	foxo	Forkhead box protein O	36 (31.0%)	14 (20.3%)	no
	5	FBpp0081986	Hsp70Aa	Major heat shock 70 kDa protein Aa	23 (19.8%)	4 (5.8%)	no
	6	FBpp0070899	schlank	Schlank, isoform A	31 (26.7%)	12 (17.4%)	no
	7	FBpp0086226	Sod2	Superoxide dismutase [Mn], mitochondrial	31 (26.7%)	8 (11.6%)	no
	8	FBpp0088134	CaMKI	Calmodulin-dependent protein kinase activity	29 (25.0%)	10 (14.5%)	no
	9	FBpp0077974	park	E3 ubiquitin-protein ligase parkin	20 (17.2%)	1 (1.4%)	no
	10	FBpp0305095	Hsp83	Heat shock protein 83	26 (22.4%)	10 (14.5%)	yes
Mouse	1	ENSMUSP00000056668	Igf-1	Insulin-like growth factor 1	10 (19.6%)	16 (51.6%)	no
	2	ENSMUSP00000029175	Src	Neuronal proto-oncogene tyrosine-protein kinase Src	3 (5.9%)	12 (38.7%)	no
	3	ENSMUSP00000050683	Foxo3	Forkhead box protein O3	8 (15.7%)	13 (41.9%)	no
	4	ENSMUSP00000055308	Foxo1	Forkhead box protein O1	5 (9.8%)	11 (35.5%)	no
	5	ENSMUSP00000000369	Rem1	GTP-binding protein REM 1	7 (13.7%)	10 (32.3%)	yes
	6	ENSMUSP00000101315	Pik3cd	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit delta isoform	1 (2.0%)	7 (22.6%)	no
	7	ENSMUSP00000102538	Ngf	Beta-nerve growth factor	5 (9.8%)	8 (25.8%)	yes
	8	ENSMUSP00000031697	Cul1	Cullin-1	6 (11.8%)	7 (22.6%)	no
	9	ENSMUSP00000120152	Stat3	Signal transducer and activator of transcription 3	11 (21.6%)	14 (45.2%)	no
	10	ENSMUSP00000115578	Ubc	Polyubiquitin-C	15 (29.4%)	4 (12.9%)	no
Yeast	1	YLR167W	RPS31	Fusion-protein cleaved to yield ribosomal protein S31 and ubiquitin	0 (0.0%)	58 (17.3%)	no
	2	YIL133C	RPL16A	Ribosomal 60S subunit protein L16A	0 (0.0%)	51 (15.2%)	no
	3	YBR048W	RPS11B	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	52 (15.5%)	no
	4	YPL090C	RPS6A	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	51 (15.2%)	no
	5	YGL103W	RPL28	Ribosomal 60S subunit protein L28	0 (0.0%)	61 (18.2%)	no
	6	YNL096C	RPS7B	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	50 (14.9%)	no
	7	YJR145C	RPS4A	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	51 (15.2%)	no
	8	YNL069C	RPL16B	Ribosomal 60S subunit protein L16B	0 (0.0%)	48 (14.3%)	no
	9	YBR031W	RPL4A	Ribosomal 60S subunit protein L4A	1 (2.2%)	53 (15.8%)	no
	10	YNL302C	RPS19B	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	47 (14.0%)	no

that the feature value "interaction with Hsp83 = yes" is more associated with the pro-longevity class than "interaction with Hsp83 = no". Actually, among the genes/proteins (instances) in our dataset that interact with Hsp83, 72.2% are pro-longevity genes, whilst among the genes/proteins that do not interact with Hsp83, 60.4% are pro-longevity genes. So, interacting with Hsp83 is associated with an increased probability of the pro-longevity class label.

However, when we look at the data partitioned by the values of the "interaction with Hsc70-4" feature in the second and third rows of the table, a different pattern emerges. Among genes/proteins that do not interact with Hsc70-4, the relative frequency of the pro-longevity class is higher among genes/proteins that do not interact with Hsp83 (58.4%) than among genes/proteins interacting with Hsp83 (50%). The same pattern is observed among genes/proteins interacting with Hsc70-4 (83.3% for Hsp83=no vs 76.7% for Hsp83=yes). Hence, in both sub-groups of genes/proteins (interacting or not with Hsc70-4), interacting with Hsp83 is associated with decreased probability of the pro-longevity class label, the reverse of the direction of association observed for the aggregated data.

Tables 7 and 8 indicate Simpson's paradox occurrences for the feature Hsp83 in the fly dataset and features Rem1 and Ngf in the mouse dataset. Hence, when interpreting the association between those features and the class variable, one should be aware of those paradox occurrences to avoid drawing wrong conclusions about the data.

4 Conclusion

This work addresses classification with uncertain categorical features, whose values are represented by probability distributions. We have proposed two new ensemble approaches called Biased Bootstrap (BB) and Biased Splitting (BS) for coping with this type of uncertainty, based on the principle that features with lower uncertainty degrees have better class-discrimination potential since there is higher confidence on their actual values across the dataset.

Our experiments have evaluated these two approaches on 10 datasets in the domains of ageing-related genes

This article has been accepted for publication in IEEE/ACM Transactions on Computational Biology and Bioinformatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2022.3218588

TABLE 8
Top-10 PPI Features in the ENB-EV+BRS Model According to the Importance Measure Based on Sufficiency

Organism	Feat. Rank	STRING ID	Gene Symbol	Protein Name	Freq. in Pro-long.	Freq. in Anti-long.	Parad.
Worm	1	F42G8.12	isp-1	Cytochrome b-c1 complex subunit Rieske, mitochondrial	9 (3.5%)	58 (11.5%)	no
	2	C26F1.4.2	rps-30	40S ribosomal protein S30	5 (1.9%)	43 (8.5%)	no
	3	C34E10.6.1	atp-2	ATP synthase subunit beta, mitochondrial	6 (2.3%)	58 (11.5%)	no
	4	B0250.1	rpl-2	60S ribosomal protein L8	3 (1.2%)	44 (8.7%)	no
	5	F40F11.1.2	rps-11	Ribosomal protein, small subunit	4 (1.6%)	53 (10.5%)	no
	6	B0393.1.1	rps-0	40S ribosomal protein SA	6 (2.3%)	51 (10.1%)	no
	7	T05E11.1	rps-5	40S ribosomal protein S5	5 (1.9%)	60 (11.9%)	no
	8	F28D1.7.1	rps-23	40S ribosomal protein S23	4 (1.6%)	48 (9.5%)	no
	9	C49H3.11.1	rps-2	40S ribosomal protein S2	7 (2.7%)	57 (11.3%)	no
	10	Y56A3A.19	Y56A3A.19	Acyl carrier protein	2 (0.8%)	51 (10.1%)	no
Fly	1	FBpp0305736	Sod	Superoxide dismutase [Cu-Zn]	30 (25.9%)	4 (5.8%)	no
	2	FBpp0081956	Hsp70Ab	Heat shock protein 70Ab	22 (19.0%)	4 (5.8%)	no
	3	FBpp0082516	Hsc70-4	Heat shock 70 kDa protein cognate 4	34 (29.3%)	9 (13.0%)	no
	4	FBpp0081986	Hsp70Aa	Major heat shock 70 kDa protein Aa	23 (19.8%)	4 (5.8%)	no
	5	FBpp0086226	Sod2	Superoxide dismutase [Mn], mitochondrial	31 (26.7%)	8 (11.6%)	no
	6	FBpp0293589	foxo	Forkhead box protein O	36 (31.0%)	14 (20.3%)	no
	7	FBpp0077974	park	E3 ubiquitin-protein ligase parkin	20 (17.2%)	1 (1.4%)	no
	8	FBpp0070899	schlank	Schlank, isoform A	31 (26.7%)	12 (17.4%)	no
	9	FBpp0088134	CaMKI	Calmodulin-dependent protein kinase activity	29 (25.0%)	10 (14.5%)	no
	10	FBpp0078604	Aux	Auxilin, isoform A	17 (14.7%)	3 (4.3%)	no
Mouse	1	ENSMUSP00000056668	Igf-1	Insulin-like growth factor 1	10 (19.6%)	16 (51.6%)	no
	2	ENSMUSP00000029175	Src	Neuronal proto-oncogene tyrosine-protein kinase Src	3 (5.9%)	12 (38.7%)	no
	3	ENSMUSP00000050683	Foxo3	Forkhead box protein O3	8 (15.7%)	13 (41.9%)	no
	4	ENSMUSP00000055308	Foxo1	Forkhead box protein O1	5 (9.8%)	11 (35.5%)	no
	5	ENSMUSP00000101553	Ins2	Insulin-2	6 (11.8%)	13 (41.9%)	no
	6	ENSMUSP00000099878	Rps6	40S ribosomal protein S6	3 (5.9%)	8 (25.8%)	no
	7	ENSMUSP00000021090	Grb2	Growth factor receptor-bound protein 2	3 (5.9%)	8 (25.8%)	no
	8	ENSMUSP00000099621	Rpa2	Replication protein A 32 kDa subunit	12 (23.5%)	1 (3.2%)	no
	9	ENSMUSP00000120152	Stat3	Signal transducer and activator of transcription 3	11 (21.6%)	14 (45.2%)	no
	10	ENSMUSP00000102538	Ngf	Beta-nerve growth factor	5 (9.8%)	8 (25.8%)	yes
Yeast	1	YLR167W	RPS31	Fusion-protein cleaved to yield ribosomal protein S31 and ubiquitin	0 (0.0%)	58 (17.3%)	no
	2	YNL096C	RPS7B	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	50 (14.9%)	no
	3	YJR145C	RPS4A	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	51 (15.2%)	no
	4	YGL103W	RPL28	Ribosomal 60S subunit protein L28	0 (0.0%)	61 (18.2%)	no
	5	YBR048W	RPS11B	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	52 (15.5%)	no
	6	YKR094C	RPL40B	Ubiquitin-ribosomal 60S subunit protein L40B fusion protein	0 (0.0%)	59 (17.6%)	no
	7	YIL133C	RPL16A	Ribosomal 60S subunit protein L16A	0 (0.0%)	51 (15.2%)	no
	8	YGL030W	RPL30	Ribosomal 60S subunit protein L30	0 (0.0%)	50 (14.9%)	no
	9	YGL123W	RPS2	Protein component of the small (40S) subunit	1 (2.2%)	58 (17.3%)	no
	10	YKL009W	MRT4	Protein involved in mRNA turnover and ribosome assembly	0 (0.0%)	55 (16.4%)	no

and drugs' side effects. For this evaluation, we have used real data with uncertain features referring to probabilities of protein-protein and protein-chemical interactions. Our results show that the ensembles using the proposed BB and BS approaches achieved higher predictive performance than baseline methods without uncertainty-based random sampling bias, with the caveat that the BB approach did not produce good results when applied to NB ensembles. These results support the hypothesis that the proposed approaches can effectively cope with uncertainty in categorical features. In particular, our proposed approaches improved the performance of RF-DFE, a Random Forest method that applies the most sophisticated approach in the literature to handle uncertain data in ensembles of this kind.

Furthermore, we have proposed two new approaches for interpreting an ensemble of Naive Bayes classifiers based on feature importance measures used to rank features in decreasing order of their influence in the ensemble's predictions. The first approach is straightforwardly based on conditional probability differences, while the second, more sophisticated, is based on the concept of a minimal set of sufficient features for classifying each instance. We have applied these two feature-ranking approaches to the ageing-related datasets and compared them to the feature ranking produced with a conventional feature importance measure for random forests. An analysis of the top-ranked features showed that, overall, they fit the current knowledge about the influence of genes/proteins on ageing well. Besides, we have also pointed out some strong patterns involving longevity effects and genes that are not included in GenAge [8]. These findings suggest some targets for future biological experiments that could confirm the longevity effects of some genes/proteins.

ACKNOWLEDGMENTS

This study was financed in part by: CNPq (Brazil) grant number 315750/2021-9; CAPES (Brazil) [finance code 001];

TABLE 9
Top-10 PPI Features in the RF-DFE+BB+BS Model

Organism	Feat. Rank	STRING ID	Gene Symbol	Protein Name	Freq. in Pro-long.	Freq. in Anti-long.	Parad.
Worm	1	M7.5	atg-7	AuTophaGy (Yeast Atg homolog)	33 (12.8%)	18 (3.6%)	no
	2	F40F11.1.2	rps-11	Ribosomal protein, small subunit	4 (1.6%)	53 (10.5%)	no
	3	Y37D8A.14	cco-2	Cytochrome c oxidase subunit 5A, mitochondrial	2 (0.8%)	48 (9.5%)	no
	4	B0412.4	rps-29	Ribosomal protein, small subunit	4 (1.6%)	46 (9.1%)	no
	5	Y45G12B.1a	nuo-5	NADH Ubiquinone Oxidoreductase	2 (0.8%)	41 (8.1%)	no
	6	F42C5.8	rps-8	40S ribosomal protein S8	4 (1.6%)	49 (9.7%)	no
	7	Y37E3.8a	Y37E3.8	Protein Y37E3.8, isoform a (Y37E3.8) mRNA, complete cds	3 (1.2%)	46 (9.1%)	no
	8	Y57G11C.34	mrps-7	28S ribosomal protein S7, mitochondrial	2 (0.8%)	54 (10.7%)	no
	9	Y105E8A.16.1	rps-20	Ribosomal protein, small subunit	3 (1.2%)	48 (9.5%)	no
	10	F58F12.1	F58F12.1	ATP synthase subunit delta, mitochondrial	4 (1.6%)	58 (11.5%)	no
Fly	1	FBpp0085780	CG15116	Glutathione peroxidase activity	21 (18.1%)	3 (4.3%)	no
	2	FBpp0070416	ph-p	Polyhomeotic-proximal chromatin protein	3 (2.6%)	4 (5.8%)	no
	3	FBpp0071973	Pi3K59F	Phosphotidylinositol 3 kinase 59F, a.k.a. Vacuolar protein sorting 34	21 (18.1%)	4 (5.8%)	no
	4	FBpp0070717	dhd	Thioredoxin-1	10 (8.6%)	1 (1.4%)	no
	5	FBpp0078138	CG7133	annotation not available	2 (1.7%)	0 (0.0%)	no
	6	FBpp0072932	PHGPx	Peroxidase activity	19 (16.4%)	2 (2.9%)	no
	7	FBpp0083975	Atg6	Beclin-1-like protein; Autophagy-related 6	23 (19.8%)	4 (5.8%)	no
	8	FBpp0082927	Prx3	Thioredoxin peroxidase 3	10 (8.6%)	0 (0.0%)	no
	9	FBpp0087354	Prx2540-2	Peroxiredoxin 2540-2	8 (6.9%)	0 (0.0%)	no
	10	FBpp0099922	Nos	Nitric oxide synthase	5 (4.3%)	0 (0.0%)	no
Mouse	1	ENSMUSP00000128260	Tfdp2	Transcription factor dp2	1 (2.0%)	1 (3.2%)	no
	2	ENSMUSP00000126874	Ccnt1	Cyclin-T1	1 (2.0%)	2 (6.5%)	no
	3	ENSMUSP00000101315	Pik3cd	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit delta isoform	1 (2.0%)	7 (22.6%)	no
	4	ENSMUSP00000030464	Pik3r3	Phosphatidylinositol 3-kinase regulatory subunit gamma	1 (2.0%)	7 (22.6%)	no
	5	ENSMUSP00000099991	Pdpk1	3-phosphoinositide-dependent protein kinase 1	1 (2.0%)	6 (19.4%)	no
	6	ENSMUSP00000021090	Grb2	Growth factor receptor-bound protein 2	3 (5.9%)	8 (25.8%)	no
	7	ENSMUSP00000025749	Rps6kb2	Ribosomal protein S6 kinase beta-2	1 (2.0%)	4 (12.9%)	no
	8	ENSMUSP00000038514	Irs2	Insulin receptor substrate 2	4 (7.8%)	9 (29.0%)	no
	9	ENSMUSP00000034296	Pik3r2	Phosphatidylinositol 3-kinase regulatory subunit beta	3 (5.9%)	8 (25.8%)	no
	10	ENSMUSP00000047839	Ppp1r13l	RelA-associated inhibitor	3 (5.9%)	0 (0.0%)	no
Yeast	1	YKR094C	RPL40B	Ubiquitin-ribosomal 60S subunit protein L40B fusion protein	0 (0.0%)	59 (17.6%)	no
	2	YKL148C	SDH1	Flavoprotein subunit of succinate dehydrogenase	11 (23.9%)	12 (3.6%)	no
	3	YKL180W	RPL17A	Ribosomal 60S subunit protein L17A	0 (0.0%)	56 (16.7%)	no
	4	YBR143C	SUP45	Polypeptide release factor (eRF1) in translation termination	0 (0.0%)	53 (15.8%)	no
	5	YER056C-A	RPL34A	Ribosomal 60S subunit protein L34A	0 (0.0%)	52 (15.5%)	no
	6	YBR048W	RPS11B	Protein component of the small (40S) ribosomal subunit	0 (0.0%)	52 (15.5%)	no
	7	YGR148C	RPL24B	Ribosomal 60S subunit protein L24B	0 (0.0%)	48 (14.3%)	no
	8	YBL092W	RPL32	Ribosomal 60S subunit protein L32	0 (0.0%)	50 (14.9%)	no
	9	YOR065W	CYT1	Cytochrome c1, heme protein, mitochondrial	9 (19.6%)	11 (3.3%)	no
	10	YLR009W	RLP24	Essential protein required for ribosomal large subunit biogenesis	0 (0.0%)	52 (15.5%)	no

TABLE 10
Simpson's Paradox Occurrence in the Fly Dataset

	Hsp83 = no		Hsp83 = yes	
	Total	Pro-long.	Total	Pro-long.
Aggregated data	149	90 (60.4%)	36	26 (72.2%)
Hsc70-4 = no	137	80 (58.4%)	6	3 (50.0%)
Hsc70-4 = yes	12	10 (83.3%)	30	23 (76.7%)

FAPERJ (Brazil) grant number E-26/201.139/2022; and Instituto Brasileiro de Geografia e Estatística (IBGE, Brazil).

REFERENCES

[1] F. Angiulli and F. Fassetti, "Nearest neighbor-based classification of uncertain data," *ACM Trans. Knowl. Discov. Data*, vol. 7, no. 1, 2013.

[2] J. Ge, Y. Xia, and C. Nadungodage, "UNN: A neural network for uncertain data classification," in *Advances in Knowledge Discovery and Data Mining*, M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi, Eds. Berlin: Springer, 2010, pp. 449–460.

[3] M. R. H. Maia, A. Plastino, and A. A. Freitas, "An ensemble of naive bayes classifiers for uncertain categorical data," in *Proc. of the 2021 IEEE Int. Conf. on Data Mining*, 2021, pp. 1222–1227.

[4] S. Tsang, B. Kao, K. Y. Yip, W.-S. Ho, and S. D. Lee, "Decision trees for uncertain data," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 1, pp. 64–78, 2011.

[5] Z. Xie, Y. Xu, and Q. Hu, "Uncertain data classification with additive kernel support vector machine," *Data Knowl. Eng.*, vol. 117, pp. 87–97, 2018.

[6] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[7] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[8] R. Tacutu, D. Thornton, E. Johnson, A. Budovsky, D. Barardo, T. Craig, E. Diana, G. Lehmann, D. Toren, J. Wang, V. E. Fraifeld, and J. P. de Magalhães, "Human Ageing Genomic Resources: new and updated databases," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1083–D1090, 2017.

[9] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. Mering, "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, 2018.

[10] M. Kuhn, I. Letunic, L. J. Jensen, and P. Bork, "The SIDER database of drugs and side effects," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1075–D1079, 2015.

[11] D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork, and M. Kuhn, "STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D380–D384, 2015.

[12] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms*. Cambridge: Cambridge University Press, 2011.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, 2018.

[14] C. B. Azodi, J. Tang, and S.-H. Shiu, "Opening the black box: Interpretable machine learning for geneticists," *Trends Genet.*, vol. 36, no. 6, pp. 442–455, 2020.

[15] A. Freitas, D. Wieser, and R. Apweiler, "On the importance of comprehensible classification models for protein function prediction," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 7, no. 1, pp. 172–182, 2010.

[16] G. Ridgeway, D. Madigan, T. Richardson, and J. O'Kane, "Interpretable boosted naïve bayes classification," in *Proc. of the Fourth Int. Conf. on Knowledge Discovery and Data Mining*, ser. KDD'98. AAAI Press, 1998, p. 101–104.

[17] T. Mori and N. Uchihira, "Balancing the trade-off between accuracy and interpretability in software defect prediction," *Empir. Softw. Eng.*, vol. 24, no. 2, pp. 779–825, 2019.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," *in Proc. of the AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.

[19] D. S. Watson, L. Gultchin, A. Taly, and L. Floridi, "Local explanations via necessity and sufficiency: Unifying theory and practice," in *Proc. of the 37th Conf. on Uncertainty in Artificial Intelligence*, 2021.

[20] B. Qin, Y. Xia, and F. Li, "DTU: A decision tree for uncertain data," in *Advances in Knowledge Discovery and Data Mining*, T. Theeramunkong, B. Kijsirikul, N. Cercone, and T.-B. Ho, Eds. Berlin: Springer, 2009, pp. 4–15.

[21] H. Boström and U. Norinder, "Utilizing information on uncertainty for in silico modeling using random forests," in *Proc. of the 3rd Skövde Workshop on Information Fusion Topics*, 2009, pp. 59–62.

[22] U. Norinder and H. Boström, "Introducing uncertainty in predictive modeling—friend or foe?" *J. Chem. Inf. Model.*, vol. 52, no. 11, pp. 2815–2822, 2012.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[24] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 2000.

[25] ——, "Comment: Understanding Simpson's paradox," *Am. Stat.*, vol. 68, no. 1, pp. 8–13, 2014.

[26] G. Shmueli and I. Iahav, "The forest or the trees? Tackling Simpson's paradox with classification trees," *Prod. Oper. Manag.*, vol. 27, no. 4, pp. 696–716, 2018.

[27] S. Nembrini, I. R. König, and M. N. Wright, "The revival of the Gini importance?" *Bioinformatics*, vol. 34, no. 21, pp. 3711–3718, 2018.

[28] X. Sun, W.-D. Chen, and Y.-D. Wang, "Daf-16/foxo transcription factor in aging and longevity," *Front. Pharmacol.*, vol. 8, p. 548, 2017.

[29] Y. Gonskikh and N. Polacek, "Alterations of the translation apparatus during aging and stress response," *Mech. Ageing Dev.*, vol. 168, pp. 30–36, 2017.

[30] N. Sun, R. J. Youle, and T. Finkel, "The mitochondrial basis of aging," *Mol. Cell*, vol. 61, no. 5, pp. 654–666, 2016.

[31] M. Fernandes, C. Wan, R. Tacutu, D. Barardo, A. Rajput, J. Wang, H. Thoppil, D. Thornton, C. Yang, A. Freitas, and J. P. de Magalhães, "Systematic analysis of the gerontome reveals links between aging and age-related diseases," *Hum. Mol. Genet.*, vol. 25, no. 21, pp. 4804–4818, 2016.

[32] P. Verbeke, J. Fonager, B. F. C. Clark, and S. I. S. Rattan, "Heat shock response and ageing: Mechanisms and applications," *Cell Biol. Int.*, vol. 25, no. 9, pp. 845–857, 2001.

[33] F. L. Muller, M. S. Lustgarten, Y. Jang, A. Richardson, and H. Van Remmen, "Trends in oxidative aging theories," *Free Radic. Biol. Medicine*, vol. 43, no. 4, pp. 477–503, 2007.

[34] J. P. de Magalhães and A. Matsuda, "Genome-wide patterns of genetic distances reveal candidate loci contributing to human population-specific traits," *Ann. Hum. Genet.*, vol. 76, no. 2, pp. 142–158, 2012.

[35] C. J. Kenyon, "The genetics of ageing," *Nature*, vol. 464, no. 7288, pp. 504–512, 2010.

Marcelo Rodrigues de Holanda Maia is a systems analyst at Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro, Brazil. He received the BSc, MSc and DSc degrees in computer science in 2008, 2015 and 2022, respectively, from Universidade Federal Fluminense, Niterói, Brazil. He undertook a research visit to the University of Kent, Canterbury, UK, from 2020 to 2021. His current research interests concentrate on data science and combinatorial optimization.

Alexandre Plastino is a full professor at Universidade Federal Fluminense, Niterói, Brazil. He obtained his BSc (1988) and MSc (1990) degrees in computer science from Universidade Federal do Rio de Janeiro and his DSc degree in computer science from Pontifícia Universidade Católica do Rio de Janeiro (2000). In 2008, he conducted a postdoctoral research at the University of Kent, Canterbury, UK. His current research interests concentrate on data science and combinatorial optimization.

Alex A. Freitas is a professor of computational intelligence at the University of Kent, Canterbury, UK. He has an interdisciplinary academic background, with a PhD degree in computer science (University of Essex, UK, 1997), in the area of machine learning (or data mining); and a research-oriented master's degree (an MPhil) in Biological Sciences (University of Liverpool, UK, 2011), in the area of the biology of ageing. His main research interests are machine learning and the biology of ageing.

João Pedro de Magalhães graduated (1999) in microbiology from Escola Superior de Biotecnologia, Porto, Portugal, and obtained his PhD (2004) from the University of Namur, Belgium. Following a postdoc at Harvard Medical School, in 2008 Prof de Magalhães joined the University of Liverpool and, in 2022, he was recruited to the University of Birmingham where he leads the Genomics of Ageing and Rejuvenation Lab (http://rejuvenomicslab.com/). The group's research broadly focuses on understanding the genetic, cellular, and molecular mechanisms of ageing.