Computer Methods and Programs in Biomedicine 244 (2024) 107934

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

Novel ensemble feature selection techniques applied to high-grade gastroenteropancreatic neuroendocrine neoplasms for the prediction of survival

Anna Jenul [a,*,1], Henning Langen Stokmo [b,c,1], Stefan Schrunner [a], Geir Olav Hjortland [d], Mona-Elisabeth Revheim [b,c,e], Oliver Tomic [a]

[a] Department of Data Science, Norwegian University of Life Sciences, Universitetstunet 3, 1433 Ås, Norway
[b] Department of Nuclear Medicine, Division of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo, Norway
[c] Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[d] Department of Oncology, Oslo University Hospital, Oslo, Norway
[e] The Intervention Centre, Division of Technology and Innovation, Oslo University Hospital, Oslo, Norway

ARTICLE INFO

Keywords:
Feature selection
Survival analysis
Expert- and data-driven model
Neuroendocrine neoplasia

ABSTRACT

Background and Objective: Determining the most informative features for predicting the overall survival of patients diagnosed with high-grade gastroenteropancreatic neuroendocrine neoplasms is crucial to improve individual treatment plans for patients, as well as the biological understanding of the disease. The main objective of this study is to evaluate the use of modern ensemble feature selection techniques for this purpose with respect to (a) quantitative performance measures such as predictive performance, (b) clinical interpretability, and (c) the effect of integrating prior expert knowledge.

Methods: The Repeated Elastic Net Technique for Feature Selection (RENT) and the User-Guided Bayesian Framework for Feature Selection (UBayFS) are recently developed ensemble feature selectors investigated in this work. Both allow the user to identify informative features in datasets with low sample sizes and focus on model interpretability. While RENT is purely data-driven, UBayFS can integrate expert knowledge a priori in the feature selection process. In this work, we compare both feature selectors on a dataset comprising 63 patients and 110 features from multiple sources, including baseline patient characteristics, baseline blood values, tumor histology, imaging, and treatment information.

Results: Our experiments involve data-driven and expert-driven setups, as well as combinations of both. In a five-fold cross-validated experiment without expert knowledge, our results demonstrate that both feature selectors allow accurate predictions: A reduction from 110 to approximately 20 features (around 82%) delivers near-optimal predictive performances with minor variations according to the choice of the feature selector, the predictive model, and the fold. Thereafter, we use findings from clinical literature as a source of expert knowledge. In addition, expert knowledge has a stabilizing effect on the feature set (an increase in stability of approximately 40%), while the impact on predictive performance is limited.

Conclusions: The features *WHO Performance Status*, *Albumin*, *Platelets*, *Ki-67*, *Tumor Morphology*, *Total MTV*, *Total TLG*, and SUV_{max} are the most stable and predictive features in our study. Overall, this study demonstrated the practical value of feature selection in medical applications not only to improve quantitative performance but also to deliver potentially new insights to experts.

1. Introduction

In the last decade, several artificial intelligence techniques have been used in either classification problems or prediction problems

[*] Corresponding author.
E-mail addresses: anna.jenul@nmbu.no (A. Jenul), h.l.stokmo@studmed.uio.no (H.L. Stokmo), stefan.schrunner@nmbu.no (S. Schrunner), goo@ous-hf.no (G.O. Hjortland), monar@ous-hf.no (M.-E. Revheim), oliver.tomic@nmbu.no (O. Tomic).
[1] These authors contributed equally.

https://doi.org/10.1016/j.cmpb.2023.107934
Received 15 June 2023; Received in revised form 5 September 2023; Accepted 17 November 2023
Available online 22 November 2023

in cancer. Types of cancers include prostate cancer, breast cancer, ear-nose-throat cancer, urological cancer, and gastrointestinal cancer, among others [1]. We can generally divide the tasks into diagnosis and staging, lesion segmentation, and prognosis and treatment response. For the prediction of survival in oral cancer, breast cancer, and lung cancer, several machine learning methods have been employed. These include artificial neural networks, decision trees, Bayesian networks, and support vector machines [2].

Gastroenteropancreatic (GEP) neuroendocrine neoplasms (NEN), in particular, are heterogeneous types of malignancies increasingly common over the last three decades [3,4]. High-grade GEP NEN encompasses both neuroendocrine tumors grade 3 (NET G3) and neuroendocrine carcinomas (NEC), where NEC is further subdivided into small cell (SC) and large cell carcinomas (LC). According to the WHO 2019 Classification of Tumors: Digestive System Tumors, NET G3 are well differentiated (WD), whilst NEC are poorly differentiated (PD), both with a Ki-67 proliferation index (Ki-67) > 20% [5]. Although both NET G3 and NEC share features of immunohistochemical staining with chromogranin A and synaptophysin, they are considered morphologically different [6].

The prognosis for patients with advanced GEP NEC is poor, with a median survival of less than 12 months [7,8], whilst the prognosis for locoregional GEP NEC is higher; 20.7 months [9]. Numerous recently published studies [7,10–16] have shown the prognostic importance of several parameters on overall survival (OS), such as age, performance status (PS), primary tumor site, tumor differentiation, TNM-stage, serum lactate dehydrogenase (LDH), serum platelet levels, proliferation marker Ki-67, maximum standardized uptake value (SUV_{max}), total metabolic tumor volume (tMTV) and total total lesion glycolysis (tTLG). Establishing more robust prognostic parameters and validating established parameters is essential to provide optimal care for this patient group.

Forecasting the OS of cancer patients as a major indicator of treatment success by machine learning models is of high relevance to offer optimal individual treatment for patients, and an active field of research [17–20]. In particular, accurate outcome prediction models pave the way for decision support in clinical practice. Since GEP NEN are rare, however, the data basis for training purely data-driven models is limited, leading to problems like overfitting, spurious correlations, and consequently to inaccurate predictions [21–23]. Two major approaches are at hand to overcome these issues: (a) increasing the number of samples (either by collecting more data or by artificial data augmentation) or (b) reducing the dimensionality of the feature space. In this work, we elaborate on approach (b), where our method of choice is feature selection. While general dimensionality reduction methods like Principal Component Analysis [24] transform the data to a new domain and thereby make identification of influencing factors difficult, feature selection reduces the dimension by subsetting the dataset by columns. As a result, a subset of the original features is retained, and the interpretability of the data columns is preserved. Using conventional feature selection methods, this approach has been successfully applied in cancer research [25,26].

Beyond the obvious benefit that predictive models become tractable, feature selection has the potential to improve the understanding of biological processes by clinical experts [27]. In particular, feature selectors point to input data parameters, which are related to explaining the target variable by a data-driven model. This information may either support or contradict existing hypotheses about the underlying biological processes or disclose previously unknown relations. Outside the context of feature selection, the importance of achieving interpretability along with high predictive performance in modeling tumor survival is further discussed in [28]. The evaluation and interpretation of the findings require close collaboration between clinical experts and data scientists. However, such an application of feature selectors is still less common in machine learning [29], where the focus typically lies exclusively on optimizing performance metrics.

Several state-of-the-art feature selection methods have been popularized in recent years, like INTERACT, CFS, InfoGain, ReliefF, and SVM-RFE [30]. State-of-the-art research in feature selection with applications in healthcare, such as L1 regularization [31], decision trees [32], Laplace scores [33], or the minimum redundancy-maximum relevance (mRMR) criterion [34], are mainly data-driven and may suffer from well-known limitations. Among these limitations is the problem that minor changes, such as the inclusion of new or removal of old samples, may have significant effects on the set of selected features — the property of feature sets to remain invariant under such changes to the dataset is referred to as feature selection stability and investigated in [35]. The usage of ensemble feature selectors, which train multiple feature selectors on subsets of the samples in a dataset, has recently been investigated extensively [36] and achieves a higher feature selection stability compared to a single feature selection run while retaining a similar predictive performance, as used, e.g. in random forest methods [37]. More recently, this fact has been exploited to introduce more stable feature selection methods such as the Repeated Elastic Net Technique for Feature Selection (RENT) [38] and the User-Guided Bayesian Framework for Feature Selection (UBayFS) [27]. Both methods are tailored towards healthcare applications, which offer a large potential with respect to the aspects discussed above. The benefit of these methods lies in (a) the fact that they represent opposite paradigms of purely data-driven versus hybrid data- and expert-driven feature selection and (b) that model interpretation and knowledge gain is their main objective rather than performance optimization. Yet, while RENT and UBayFS have shown to perform well compared to established feature selection methods [27,38] and additionally provide information on feature selection stability, these two methods have not yet been applied to purely clinical studies. This leaves the full clinical value of these methods unexplored, an issue that this study attempts to address.

As far as we know, no publications have yet explored feature selection techniques for the prediction of overall survival using machine learning incorporating PET-parameters in neuroendocrine neoplasias. In fact, there exist very few studies applying machine learning techniques in these cancers for survival prediction. [39] used eight machine learning models to predict overall survival after upper gastrointestinal surgery. However, they did not perform any feature selection prior to the application of the models. Others have used machine learning techniques for tasks other than overall survival prediction. [40] used gene data to predict subtypes of small intestinal neuroendocrine tumors using an SVM model, [41] used LASSO regression to predict the histological grade in pancreatic neuroendocrine tumors, and [42] used a LASSO Cox regression model to predict overall survival in oesophageal neuroendocrine carcinomas prior to treatment.

This paper aims to improve the understanding and insights into the OS in patients diagnosed with high-grade GEP NEN by applying recently developed ensemble feature selection techniques RENT and UBayFS. Thereby, we demonstrate that the applied feature selectors contribute to the clinical understanding of the disease by supplying information of high practical value. Our investigated dataset contains 63 patients diagnosed with high-grade GEP NEN. Our experiments compare both ensemble feature selectors in setups with and without the use of expert information. Our main goals are: (I) to determine the most informative set of features with respect to the outcome prediction task; (II) to interpret those selected features clinically, and compare with previously established features — to evaluate the first goal, we measure the quality of the selected feature set in terms of predictive performance and selection stability. Another aspect of interest is: (III) to determine the effect of integrating prior expert knowledge into the feature selection process, compared to a purely data-driven pipeline. In a similar analysis [43], the authors present a novel wrapper feature selector which can incorporate prior knowledge, as well. On a dataset with clinical and medical image data, they show that the inclusion of prior knowledge (in this case, standardized uptake value (SUV) features) improves the stability of feature selection. To this end, we discuss the feature selection results with respect to their clinical relevance and potential to improve our understanding of what influences the OS of GEP NEN patients.

Notations In the following, we denote the input data matrix by $\boldsymbol{X} \in \mathbb{R}^{m \times n}$, where m denotes the number of patients, and n denotes the number of features. Further, the target variable is denoted by $\boldsymbol{y} \in \mathbb{R}^{m}$. A feature set S is characterized by the indices, $S \subseteq \{1, \dots, n\}$. Vectors and matrices are indicated by bold letters.

2. Methods

2.1. GEP NEN dataset

Patient cohort Patients were identified from a single institutional cohort at Oslo University Hospital, also included in two multi-institutional Nordic NEC registries organized by the Nordic Neuroendocrine Tumor Group, previously described by [10]. In short, this cohort consisted of 192 patients included between January 2000 and July 2018, with GEP NEC classified according to the WHO 2010-classification [44]. In addition, all patients who had performed a fluorine-18 labeled 2-deoxy-2-fluoroglucose ([18F]FDG) positron emission tomography/computed tomography (PET/CT) within 90 days of their histological evaluation were eligible for inclusion. A hundred and seven patients did not have PET/CT performed, and two patients had no metabolic active lesions available for evaluation. Seventeen patients had more than 90 days between their biopsy and PET/CT, leaving 66 patients available for inclusion in this study.

Histological re-evaluation As described previously in [10], the histological re-evaluation was performed on both core biopsies and surgical specimens from GEP NEC primary tumors and metastases. These were re-classified according to the most recent WHO 2019-classification [5] and with regards to synaptophysin, chromogranin A, and the proliferation marker Ki-67. In this study, only the re-evaluated histology features were used, while the original histology block was discarded.

PET/CT acquisition All PET/CT scans were done according to the European Association of Nuclear Medicine (EANM) guidelines [45,46] as part of the clinical routine. The three PET scanners used were a 40-slice Siemens Biograph mCT hybrid PET/CT system (Siemens Healthineers, Erlangen, Germany), a Siemens Biograph 64, and a 64-slice General Electric (GE) Discovery 690 (GE Healthcare, Waukesha, WI, USA). Both Biograph PET/CTs were both EANM Research Ltd. (EARL)-accredited, whilst the Discovery 690 followed similar routine quality controls harmonizing with the two Biographs for cross-calibration. All acquisitions were from the vertex or skull base to mid-thighs. Before the PET acquisition, a low dose CT was acquired for anatomical information and attenuation correction. Parameters from PET were extracted using the ROI Visualisation, Evaluation, and Image Registration (ROVER) software v3.0.5 (ABX GmbH).[2]

Treatment All patients received treatment in the form of surgery, chemotherapy, or a combination of both. In total, 54 patients received the standard treatment of platinum-based chemotherapy. Patients could have surgery prior to or after [18F]FDG PET/CT. Evaluation of response to chemotherapy treatment was done with CT using the Response Evaluation Criteria in Solid Tumors (RECIST) [47].

Outcome variable Our outcome variable, or outcome target, was overall survival (OS) in months. This can be defined as the time a patient remains alive from the time of diagnosis to death of any cause; hence, it is not disease-specific. It is a reliable and easily available survival measure [48]. We can analyze such survival data, i.e., the time from diagnosis to the time of death, using the Kaplan-Meier estimator. Patients who did not experience the event during the time of the study (or during follow-up) (i.e., death) are said to be ‘censored’ [49]. Being ‘censored’ means that we do not know when this event will occur, only that it has not happened at the end of the study (or during follow-up). Across the full dataset, the empirical distribution of the outcome variable is illustrated as a histogram in Fig. 1.

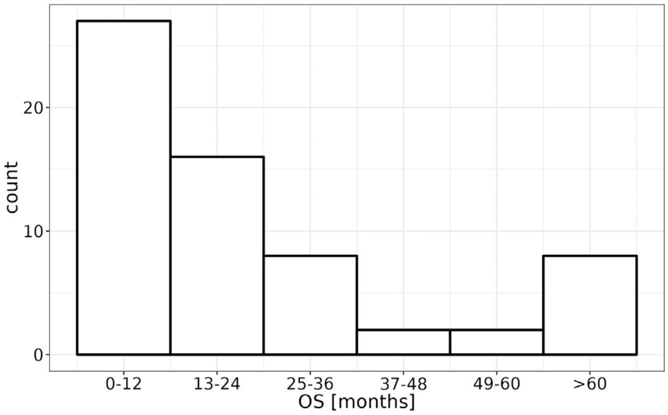


Fig. 1. Distribution of the overall survival in months.

Data blocks The data were grouped into five different blocks

(p) baseline patient characteristics
(b) baseline blood values
(h) re-evaluated histology
(i) PET/CT imaging
(t) treatment

The data contained mainly categorical and ordinal features with very few continuous variables. An overview of pairwise feature correlations[3] provided in Fig. 2.

2.2. Data preprocessing

The data preprocessing consists of several chronological steps prior to applying the ensemble feature selectors, see Fig. 3.

Data cleaning The first step in data preprocessing is to exclude features known to be unimportant, such as features with only one unique value for all patients or duplicated features. Furthermore, we remove all data columns containing more than 25% missing values across all patients. The threshold of 25% is selected as a trade-off between aiming to preserve as many features as possible, and avoiding a possible bias that may be induced by large-scale imputations. Even though a bias may already occur at a lower cutoff of 10% missing values [50], a higher value was selected to provide more flexibility to the feature selection. In the given dataset, the number of features affected by the removal, however, changes only by ±1 feature, if the threshold is decreased to 10%, or increased to 33%, respectively. By this criterion, we remove 15 features from block (p), one feature from block (b), eight features from block (h), 14 features from block (i), and eight features from block (t).

Further, three patients are excluded from the experiments due to a high number of missing values in at least one block. All subsequent

[2] The detailed imaging- and extraction protocol is described in [10].

[3] Only features with an absolute correlation > 0.5 with at least one other feature are shown.

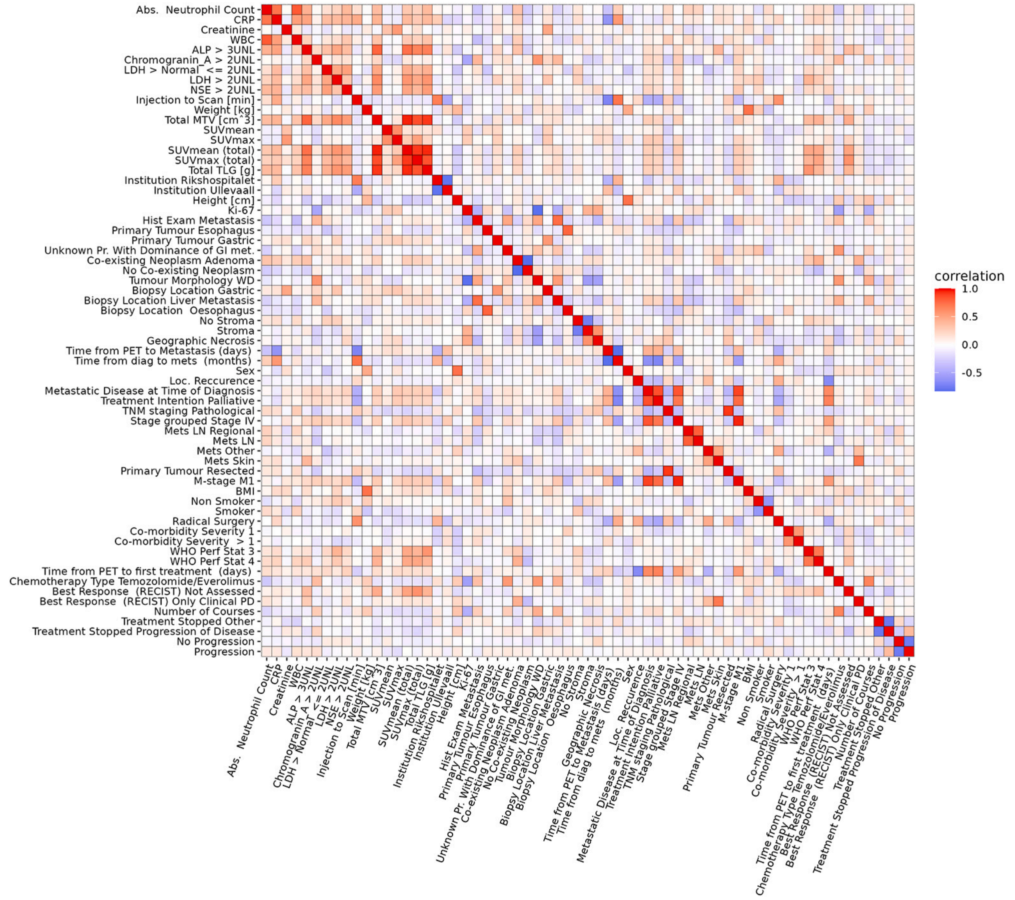


Fig. 2. Correlations between input features (features with absolute correlations ≤ 0.5 were removed).

preprocessing steps are conducted on the remaining 63 patients and are applied by block to retain the homogeneous block structure.

Missing values Some values were missing because the clinicians did not fill out the case registration forms (CRF) properly or completely. Amongst other reasons, this may be because the information was missing in the patient journal, a blood sample was not done, a parameter was forgotten registered in the patient journal, or because the patients are referred from other hospitals. Such features, which are unavailable for a large percentage of patients, cannot be assessed properly in a data-driven manner and were therefore excluded — an imputation of those features would be unreliable due to the small sample size and may introduce incorrect or misleading information into the model.

As a second step, we impute the features with less than 25% missing values via an adaptation of the k-nearest neighbors (kNN) imputation algorithm [51]. Nearest neighbor imputation strategies are frequently used for a dataset with many missing values [52]. The number of features and the number of patients that have at least one missing value for each block are: (p) (7:25), (b) (5:16), (h) (7:6), (i) (2:2), and (t) (3:3) where the first number represents the number of features and the second number represents the number of patients.

In particular, we restrict the feature space to non-missing columns and compute a matrix of pair-wise distances between all patients. We denote the set comprising the k-nearest neighbors of patient i by $N_k(i) \subseteq \{1, \dots, m\}$. Assuming that feature j is missing for patient i, we impute $x_{i,j}$ by $x_{i,j}^{\text{imp}}$, representing the median (instead of the mean, as suggested by [51]) of feature j across the patient's k nearest neighbors where the feature value is known, i.e.

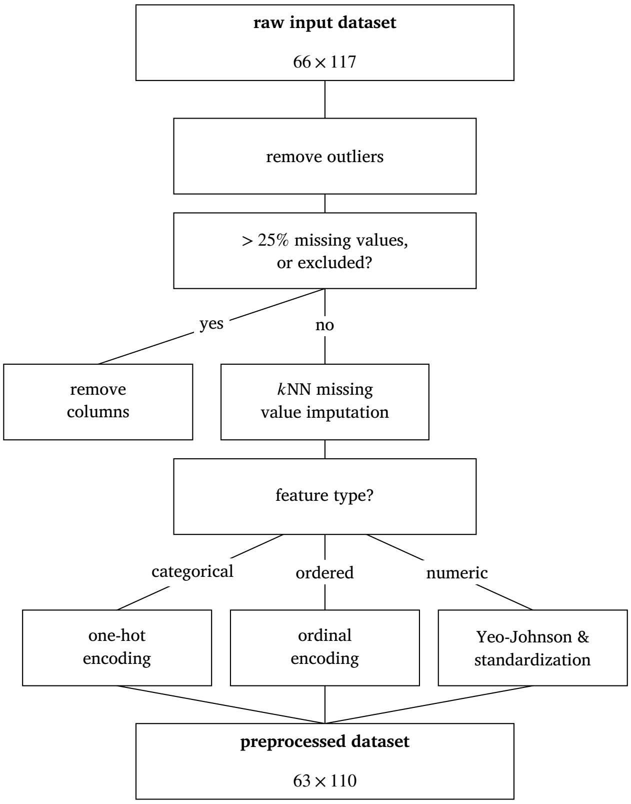


Fig. 3. Preprocessing pipeline for the dataset.

Table 1
One-hot versus ordinal encoding of a 4-level variable (levels A, B, C, D). Ordinal encoding assumes an order of the levels (here: A<B<C<D).

Level	One-hot encoding	Ordinal encoding
A	(0,0,0)	(0,0,0)
B	(0,0,1)	(0,0,1)
C	(0,1,0)	(0,1,1)
D	(1,0,0)	(1,1,1)

$$x_{i,j}^{\text{imp}} \leftarrow \text{median}\left\{x_{l,j} : l \in N_k(i)\right\}. \tag{1}$$

Ordered categorical features are transformed to an integer scale before interpolation. The usage of an odd value of k (by default, we use $k = 5$) guarantees that the median returns an integer, which is a clear benefit over the mean when using the technique for ordered features.

Categorical feature encoding One challenge in clinical data science is handling different data types [53]. Categorical features require encoding in order to be processed alongside numeric variables in predictive models. In particular, we distinguish between ordinal and nominal categorical variables: Nominal variables (i.e. variables without an internal order of the feature levels), such as *clinical institution*, are one-hot encoded [54]. Given a feature j with c_j feature levels, the one-hot encoding produces a set of $c_j - 1$ binary features $\{e_2, \dots, e_{c_j}\}$, given as follows:

$$\left(e_l\right)_i = \begin{cases} 1 & \text{if } x_{i,j} = l, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

for $l \in \{2, \dots, c_j\}$ indicating the feature level. The number of one-hot/ordinal categorical features is: 21/5 for block (p), 0/3 for block (b), 10/2 for block (h), 1/0 for block (i), and 5/0 for block (t). To avoid linear dependencies between features, the first feature level is not represented by a binary vector in the encoded space, but rather contributes to the model intercept, see Table 1.

Features with an internal order among their levels (ordinal variables), such as the *WHO performance status* with levels *0, 1, 2, 3,* and *4*, require an ordinal encoding to retain the relevant information about the order. Under the assumption that the influence of a feature increases from lower to higher levels (i.e., higher levels comprise the lower levels and an additive effect), the following encoding is used:

$$\left(e_l\right)_i = \begin{cases} 1 & \text{if } x_{i,j} \leq l, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

for feature level $l \in \{2, \dots, c_j\}$. Again, the first feature level, which would be assigned a value of 1 across all samples in the encoded space, is not assigned a binary vector in the encoded space. A comparison between one-hot and ordinal encoding is provided in Table 1. In contrast to transforming to an integer scale, this binary ordinal encoding preserves the order among the categories but does not pretend equal distances between the categories on a numerical scale.

Feature transformation and normalization During our experiments, we split the dataset into train and test sets. To normalize the distribution of each numeric feature, we use the Yeo-Johnson power transformation along with standardization [55]. The Yeo-Johnson power transformation is an extension of the well-established Box-Cox transformation with the benefit that it enables the transformation of negative and zero values. The intention is to bring the data closer to a normal distribution by simultaneously stabilizing data variance. For a given feature j, Yeo-Johnson's power transform is defined as

Table 2
Encoding of the target variable "overall survival" (OS) [months].

Level	Encoding
$\text{OS} \leq 12$	1
$12 < \text{OS} \leq 24$	2
$24 < \text{OS} \leq 36$	3
$36 < \text{OS} \leq 48$	4
$48 < \text{OS} \leq 60$	5
$60 < \text{OS}$	6

$$x_{i,j}^{\text{YJ}} \leftarrow \begin{cases} \dfrac{((x_{i,j}+1)^{\lambda_j} - 1)}{\lambda_j} & \text{if } \lambda_j \neq 0, x_{i,j} \geq 0 \\ \log(x_{i,j}+1) & \text{if } \lambda_j = 0, x_{i,j} \geq 0 \\ -\dfrac{((-x_{i,j}+1)^{2-\lambda_j} - 1)}{2-\lambda_j} & \text{if } \lambda_j \neq 2, x_{i,j} < 0 \\ -\log(-x_{i,j}+1) & \text{if } \lambda_j = 2, x_{i,j} < 0. \end{cases} \tag{4}$$

Commonly, the transformation parameter λ_j is estimated from the data using a maximum likelihood approach. After the Yeo-Johnson transformation, we scale the data to zero mean and variance of 1. To prevent biased train and test data, the transformation parameter λ_j and the mean and variance for the standardization are estimated on the training data in each split separately.

Encoding of the target variable Even though machine learning models for censored data are evolving, most present predictive models cannot handle censored data [56]. To avoid the problem presented by censored data, we encode the OS in months into an integer value (1-6). Using 60 months median follow-up time as a reference, there are no censored patients with OS below 60 months. Considering survival on a yearly basis we use the representation of the target variable in our experiments as in Table 2. Since each level in the encoded space equals one year, predictive errors used in the remainder of this paper refer to a yearly scale.

2.3. Feature Selection Methods

In this work, we investigate two ensemble feature selection methods, which have been tailored to fit the requirements of datasets in the life science domain: the Repeated Elastic Net Technique for Feature Selection (RENT) [38] and the User-Guided Bayesian Framework for Feature Selection (UBayFS) [27]. Both methods build on the principle of (a) randomly sub-sampling the input dataset and (b) training an elementary feature selection model on each sample. The final feature set is determined by applying a meta-model on the feature sets selected by the elementary models, see Fig. 4. In the case of RENT, the elementary feature selector type is restricted to elastic net regularization [57] using logistic regression models for binary classification problems or ordinary least squares linear regression models for regression problems, while UBayFS operates on an arbitrary elementary model type.

RENT The rules to obtain a final feature set further demonstrate the distinct scopes of the methods: RENT defines three criteria τ_1, τ_2 and τ_3 for the selection of features based on the distribution of their weights across the elementary models; (I) the number of times the feature weights are non-zero (τ_1) is above a level specified by the user; (II) the alternation of the sign of the feature weights does not surpass a user-defined level (τ_2); (III) the sizes of the feature weights deviate significantly from 0 (τ_3). The hyperparameters for RENT comprise a number M of elementary models, an internal data split ratio, two parameters associated with the elastic net regularization in the elementary models (C and ℓ_1), as well as one cut-off parameter for each of the three criteria τ_1, τ_2, τ_3.

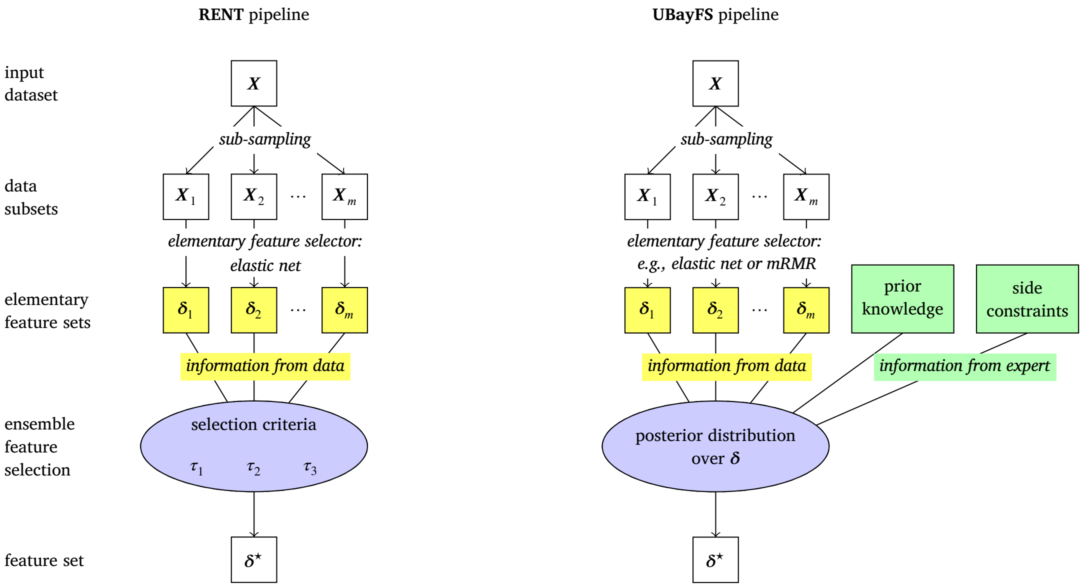


Fig. 4. Overall structure of both ensemble feature selection methods, RENT and UBayFS. After training an ensemble of m elementary feature selectors, information is combined in a meta-model. While RENT uses information from data only, UBayFS additionally includes expert information.

UBayFS In contrast, UBayFS combines the selection frequency of each feature across the elementary models with prior information from domain experts, along with side constraints. In particular, the prior weighting of features is possible, along with the definition of linear side constraints between features (and feature blocks). In practice, weights can represent knowledge about the importance of features, which is verified from previous publications. Side constraints enable the user to restrict the feature set's maximum size $\max_s$ and account for the intrinsic block structure during feature selection (e.g., in multi-source datasets). Hence, RENT implements a purely data-driven approach based on Elastic Net, while UBayFS is a general meta-model with capabilities to integrate contextual information about the data generation process. In its most basic setup, UBayFS requires as hyperparameters a number of elementary models M and an internal data split ratio, a maximum number of features $\max_s$, and a model type to use as the elementary feature selector.

2.4. Outcome prediction

Linear regression Given a set of selected features S, we make use of linear regression models [58] to model the target variable $\boldsymbol{y}$. In its simplest form, the linear regression model (with intercept) is given as

$$\boldsymbol{y} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{n+1}$ is the model parameter vector, $\tilde{\boldsymbol{X}}$ denotes the matrix containing one column of ones, followed by the sub-matrix of $\boldsymbol{X}$ restricted to the columns contained in S. Further, $\boldsymbol{\varepsilon} \underset{\text{iid}}{\sim} N(0,\sigma^2)$ denotes the model error with constant error variance $\sigma > 0$. By default, parameters of linear regression models are obtained via ordinary least squares (OLS), i.e., by minimizing the least squares error

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2. \tag{6}$$

Once the parameter vector $\boldsymbol{\beta}$ is estimated by optimizing Eq. (6) analytically, predictions are obtained by evaluating $\hat{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta}$.

k-nearest neighbor (kNN) regression As an alternative to the linear regression model, a k-nearest neighbor (kNN) regression model [58] is used to compute predictive results. In contrast to the linear regression model, the kNN model does not assume a linear relationship between the predictors and the target variable. Similar to the kNN method used for missing value imputation in Section 2.2, a neighborhood $N_k(i)$ of sample i containing the k nearest training data points with respect to a Euclidean metric on the feature space is computed for any data point $\boldsymbol{x}_i$. The prediction for the target value y_i corresponding to sample i is given by the mean of the neighbor's target values

$$\hat{y}_i = \frac{1}{k} \sum_{l \in N_k(i)} y_l. \tag{7}$$

Note that the neighborhood $N_k(i)$ is a subset of the training samples only, while $\hat{y}_i$ may represent both training or test samples.

Both predictive models, linear regression as well as the kNN regression model, are known to suffer from the curse of dimensionality — hence, we can assume that selecting a high number of features deteriorates each model. The opposite extreme for both methods, i.e., selecting no features at all, leads to predicting the output with the mean over the training data regardless of the input. Thus, we expect a well-performing feature selector to deliver a proper subset S of the feature set $\{1,\dots,n\}$, which allows both predictive models to perform better than the baselines given by (a) the overall mean of the target variable, and (b) a model including all features.

2.5. Implementation & Reproducibility

Parts of our analyses are conducted in the programming languages R [59]; other parts are conducted in Python [60]. We use the open-source implementations for RENT [61] and UBayFS [62]. For data preparation and preprocessing, we deploy the R package *caret* [63], and the Python package *scikit-learn* [64]. Fold indices are shared between R and Python. Predictive models are trained and evaluated in R using the *caret* package for all model setups. All plots are created using package *ggplot2* [65]. For reproducibility, the full code used to run and evaluate the experiments is available on the open-source software platform GitHub[4]. In addition, a detailed list of blocks and features is provided. The raw dataset, however, is restricted under data protection regulations.

All results are produced on an Intel Core i7 CPU @1.8 GHz, 32GB RAM under a Windows 11 Pro operating system.

[4] https://github.com/annajenul/GEP_NEN_Analysis.

3. Experiments and Results

Our experimental results are structured into a pre-study, where we determine optimal hyperparameters for the feature selection algorithms, followed by two main experiments. Experiment 1 focuses on the comparison of the two models, RENT and UBayFS, on the dataset without accounting for additional expert knowledge. Experiment 2 is operated on UBayFS only, as prior information and additional side constraints are included in the feature selection.

Our main focus in the experiments lies on the selected feature sets, along with the impact of the feature selection on predictive performance. We provide feature counts from both of the investigated feature selectors, RENT and UBayFS, across five different train-test splits of the dataset. Unless specified otherwise, all experiments are conducted using the hyperparameters determined during the pre-study.

3.1. Experimental setup

Model parameters Both algorithms, RENT, and UBayFS are trained on $M = 100$ ensemble models and internal 0.75/0.25-splits for subsampling the dataset. The underlying elementary feature selector for RENT is, by definition, an elastic net regularized linear regression model. Thus, RENT requires five hyperparameters to be determined during the pre-study (2 elastic net regularization parameters, ℓ_1 and C, as well as three thresholds τ_1, τ_2, and τ_3 for the selection criteria). In order to make results comparable with UBayFS, we further deploy a side condition to restrict the search space to settings, which deliver a maximum number of features $\max_s$ during validation. Thus, the number of features selected by RENT is approximately equal to the pre-defined parameter $\max_s$.

UBayFS uses minimum redundancy max relevance (mRMR) [34] as an elementary feature selector. The internal number of features in each elementary model is set to $\max_s$, i.e. each elementary model selects exactly $\max_s$ features. For the meta-model, the same parameter $\max_s$ is used to restrict the maximum number of selected features via a max-size side constraint (hard constraint) — while different levels of $\max_s$ are evaluated in experiment 1, the parameter is set to the default $\max_s = 20$ in experiment 2. Further, unless otherwise stated, prior feature weights in UBayFS are set uniformly to 0.1 across all features, which results in a non-informative prior.

Train-test splits As the ratio between the number of patients and features is unbalanced, with 63 patients and 110 encoded features, the reliability of the feature ranking results must be validated to reduce the risk of spurious correlations and overfitting [66]. Hence, we perform a 5-fold split of the dataset. For all possible permutations, we use four folds for training UBayFS or RENT, as well as the predictive models and the remaining fold for testing. Hyperparameters are determined on each split separately by internally subsetting the 4-fold training set (nested split). The 5-fold splits and hyperparameters determined in the pre-study remain the same across all experiments.

For each feature selection method, we provide the selection frequencies of each feature across the five folds, i.e., a feature obtains an importance score between 0 and 5 according to the number of folds it was selected for. For predictive performance scores, a linear regression model and a kNN regression model are trained on the same training folds, using the features from the preceding feature selection, and evaluate the prediction error on the test set (averaged across all folds).

Performance metrics To assess whether a feature set contains relevant information for training predictive models, we analyze the predictive performance in a regression setup following the feature selection step. The performance is quantified using the root mean squared error (RMSE), which has a lower bound of 0 and shall be minimized.

Using the stability criterion introduced by [35], we further evaluate the feature selection stability across the five folds for RENT and UBayFS. The computed score is asymptotically bounded in the interval $[0, 1]$; a value of 1 indicates perfect stability, i.e. the same feature set is selected in each model, while 0 indicates that selected feature sets show no overlap.

Table 3
Selected hyperparameters for each train/test split.

	Parameter	Fold				
		1	2	3	4	5
RENT	ℓ_1	0.3	0	0.3	0.3	0.3
	C	1	1	1	1	1
	τ_1	0.3	0.5	0.3	0.35	0.35
	τ_2	0.3	0.5	0.4	0.35	0.35
	τ_3	0.975	0.975	0.975	0.975	0.975
UBayFS	$\max_s$	20	20	20	20	20

Furthermore, the redundancy rate (RED) returns an intrinsic feature set quality measure by computing the average absolute Pearson correlation among the selected features. Small correlations are desirable as highly correlated features represent redundant information. Equally to the absolute Pearson correlation coefficient, RED is bounded in $[0, 1]$.

In experiment 2, we additionally assign prior weights to a subset of features — therefore, we also evaluate the percentage of prior-elevated features (PERC) in the selected feature sets as well. If PERC is high, features extracted via data-driven feature selectors match the domain experts' knowledge. However, a low PERC does not necessarily contradict expert knowledge since the features may be highly correlated, and therefore, similar information may be encoded in multiple distinct sets of features.

3.2. Pre-Study

The pre-study aims to determine the optimal hyperparameters for RENT. Given a 0.75/0.25 outer train-test split as specified above, only train data are used for hyperparameter selection. For this purpose, 4-fold cross-validation is performed on each train dataset (using the same four folds as in the outer train-test split). Across the resulting four models, hyperparameters are selected by maximizing predictive performance in a grid search over the parameter space $C \in \{1, 10, 100, 1000\}$, $\ell_1 \in \{0, 0.1, 0.2, \dots, 1\}$, $\tau_1, \tau_2 \in \{0, 0.05, 0.1, \dots, 1\}$, and $\tau_3 = 0.975$ (fixed).

The runtime for the full computation associated with the pre-study (parameter selection and final feature selection) for RENT comprised approx. 350 sec (16 cores, 24 threads in parallel). Since UBayFS does not require parameter selection, the runtime to evaluate the feature selection model for different levels of $\max_s$ (see Experiment 1) is shorter (approx. 65 sec without parallelization).

Table 3 shows the hyperparameters identified for RENT and UBayFS in each train-test split (given by the numbers of the test folds 1-5). Due to the restriction of the maximum number of features, the stated parameters may not represent global maxima for the performance of RENT; however, comparability between the methods is preserved. Furthermore, since the number of features is restricted, the selected hyperparameters are in a similar range between the folds.

3.3. Experiment 1: feature selection without prior knowledge

Having determined hyperparameters for each fold in the pre-study, RENT, and UBayFS are applied in each data split to the training dataset to select an optimized feature set for a given $\max_s$ on a purely data-driven basis.

Selected features For each feature, selection frequencies across the five test folds are further provided in Table 4 (columns *RENT* and UBayFS, $w = 0.1$).

Table 4

Feature selection frequencies across five folds by RENT and UBayFS (with different prior weight levels w for selected features). Features with increased prior weights in the UBayFS setup reported in the last column are highlighted with asterisks.

Block		Feature	RENT	UBayFS		
				$w = 0.1$	$w = 50$	$w = 110$
(p)	*	Age at Diagnosis	2	2	5	5
		Time from PET to Metastasis (days)	0	0	0	0
		Time from PET to Diagnosis (days)	0	0	0	0
		Time from diag to mets (months)	0	0	0	0
		Sex	0	0	0	0
		Loc. Adv. Resectable Disease	0	0	0	0
		Loc. Reccurence	0	0	0	0
		Metastatic Disease at Time of Diagnosis	3	0	0	0
		Treatment Intention Palliative	4	4	2	0
		Prior Other Cancer	2	2	0	0
		Living Alone	0	0	0	0
	*	TNM staging Pathological	0	0	0	0
		Stage grouped Stage IV	0	0	0	0
		Mets Bone	5	5	5	0
		Mets LN Distant	0	0	0	0
		Mets LN Regional	0	0	0	0
		Mets LN Retro	0	0	0	0
		Mets LN	0	0	0	0
		Mets Liver	0	0	0	0
		Mets Lung	0	0	0	0
		Mets Other	0	0	0	0
		Mets Skin	0	0	0	0
		Primary Tumor Resected	0	0	0	0
		M-stage M1	0	0	0	0
		BMI	1	0	0	0
		Non Smoker	0	0	0	0
		Smoker	0	0	0	0
		Radical Surgery	3	4	0	0
		Co-morbidity Severity 1	0	0	0	0
		Co-morbidity Severity > 1	0	0	0	0
		N-stage N1	0	0	0	0
		N-stage > N1	0	0	0	0
	*	WHO Perf Stat 1	0	0	4	5
	*	WHO Perf Stat 2	4	5	5	5
	*	WHO Perf Stat 3	0	0	3	4
	*	WHO Perf Stat 4	0	0	0	2
(b)		Abs. Neutrophil Count	0	0	0	0
	*	Albumin	2	5	5	5
		CRP	5	5	4	0
		Creatinine	0	0	0	0
		Haemoglobin	0	0	0	0
		WBC	1	1	1	0
		ALP > Normal <= 3UNL	4	5	3	0
		ALP > 3UNL	1	2	0	0
		Chromogranin_A > Normal <= 2UNL	0	0	0	0
		Chromogranin_A > 2UNL	0	0	0	0
	*	LDH > Normal <= 2UNL	0	0	1	5
	*	LDH > 2UNL	0	0	2	5
		NSE > Normal <= 2UNL	0	0	0	0
		NSE > 2UNL	0	0	0	0
	*	Platelets	2	5	5	5
(h)	*	Ki-67	5	5	5	5
		Hist Exam Metastasis	0	0	0	0
	*	Primary Tumor Esophagus	0	0	1	5
	*	Primary Tumor Gallbladder/duct	0	0	4	5
	*	Primary Tumor Gastric	0	0	5	5
	*	Primary Tumor Other abdominal	0	0	2	4
	*	Primary Tumor Pancreas	1	0	4	5
	*	Primary Tumor Rectum	0	0	3	5
	*	Unknown Pr. With Dominance of GI met.	0	0	0	5
		Co-existing Neoplasm Adenoma	0	0	0	0
		Co-existing Neoplasm Dysplasia	0	0	0	0
		No Co-existing Neoplasm	0	0	0	0
	*	Tumor Morphology WD	4	3	4	5
		Chromogranin A Staining	0	0	0	0
		Architecture Infiltrative	1	0	0	0
		Architecture Organoid	1	0	0	0
		Architecture Solid	0	0	0	0
		Architecture Trabecular	1	1	0	0
		Vessel Pattern Distant	1	2	0	0
		Biopsy Location Gastric	0	0	0	0
		Biopsy Location Liver Metastasis	0	0	0	0
		Biopsy Location Lymph Node	0	0	0	0
		Biopsy Location Oesophagus	0	0	0	0
		Biopsy Location Pancreas	0	0	0	0
		Biopsy Location Peritoneum	2	0	0	0
		No Stroma	4	1	0	0
		Stroma	3	3	0	0
		Geographic Necrosis	0	2	0	0
		Synaptophysin Staining 2+	0	0	0	0
		Synaptophysin Staining 3+	0	1	0	0
(i)		Injection to Scan [min]	2	2	0	0
		Weight [kg]	2	0	0	0
	*	Total MTV [cmˆ3]	3	1	5	5
		SUVmean	0	0	0	0
	*	SUVmax	2	4	5	5
		SUVmean (total)	1	0	0	0
		SUVmax (total)	5	5	5	0
	*	Total TLG [g]	4	1	5	5
		Institution Rikshospitalet	4	3	0	0
		Institution Ullevaall	0	0	0	0
		Height [cm]	0	0	0	0
		Glucose [mmol/L]	2	0	0	0
(t)		Time from PET to first treatment (days)	0	0	0	0
		Chemotherapy Type Cisplatin/Etoposide	4	3	0	0
		Chemotherapy Type Other	0	0	0	0
		Chemotherapy Type Temozolomide/Capecitabine	1	0	0	0
		Chemotherapy Type Temozolomide/Everolimus	4	5	2	0
		Best Response (RECIST) Not Assessed	0	1	0	0
		Best Response (RECIST) Only Clinical PD	0	0	0	0
		Best Response (RECIST) Partial Response	2	0	0	0
		Best Response (RECIST) Progressive Disease	0	0	0	0
		Best Response (RECIST) Stable Disease	0	0	0	0
		Reintroduction with Cisplatin Etoposide	0	0	0	0
		Number of Courses	4	4	2	0
		Treatment Stopped Other	1	2	0	0
		Treatment Stopped Progression of Disease	0	0	0	0
		Treatment Stopped Toxicity	0	0	0	0
		No Progression	5	3	2	0
		Progression	3	3	1	0

Predictive performance Further, Fig. 5 illustrates the predictive performances of kNN and linear regression models trained after UBayFS and RENT feature selection. The plot shows the RMSE for each fold given a predefined number of selected features $\max_s$.

Notably, RENT performs better using the linear regression model as the predictor, while UBayFS shows a better performance in combination with kNN. The stronger performance of RENT with linear regression may be a result of the fact that the underlying feature selection in RENT is based on a regularized linear regression model. UBayFS, however, is based on mRMR, which does not build upon a linear predictive model.

While linear regression results deteriorate at a higher number of features ($\max_s > 30$), the kNN model retains a similar performance level, which suggests that the curse of dimensionality does not yet have a strong effect on the Euclidean distance for the given feature space dimensionalities. For the linear model, overfitting is triggered by a large ratio between the number of features and the number of patients[5].

Among all compared methods, differences between the folds are obvious: for instance, fold 4 is predicted with the lowest RMSE averaged over all combinations of feature selector, predictive model, and $\max_s$. On the other hand, fold 2 is associated with the largest averaged RMSE, followed by fold 3. Potentially, differences between folds may be caused by two factors (or combinations of both):

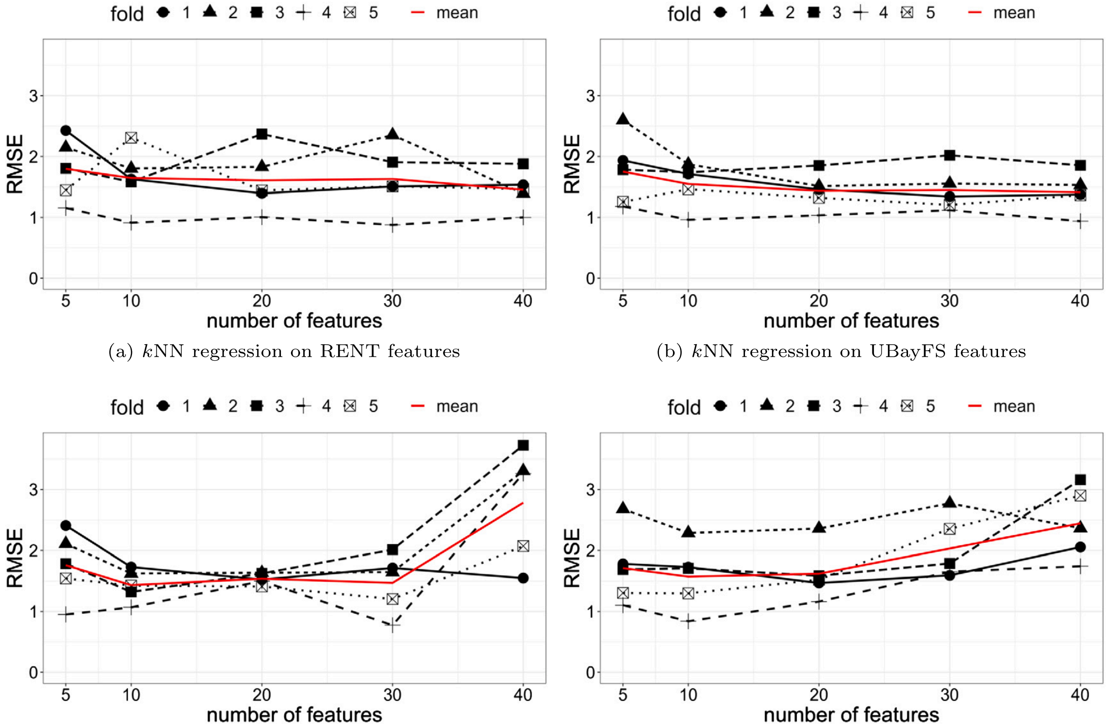
- the cohort of patients in the *training set* does not represent the global distribution of the data well — e.g., the training data do not contain a sufficient number of samples with particularly high or low target values (bad prediction due to a bad model);
- the cohort of patients in the test set is particularly hard to estimate, e.g., due to outliers (bad prediction in spite of an appropriate model);

Due to the low number of only 12-13 patients in each fold, even a low number of hard-to-predict outliers may deteriorate RMSE results significantly.

Residuals In order to shed light on the dynamics leading to the differences in performance between the data folds, histogram plots of the residuals for fold 2 (worst fold in UBayFS) and fold 4 (best fold across most setups) at $\max_s = 20$ are provided in Fig. 6. Residuals are defined as the difference between the true value and the prediction; thus, a positive or negative residual value indicates an underestimation or overestimation of the lifetime, respectively.

In contrast, to fold 4, the residuals from fold 2 are more dispersed. All histograms are symmetric and centered around 0, which indicates

[5] We also evaluated a support vector machine with radial basis function kernel. The performance was similar to the kNN model.



(a) kNN regression on RENT features

(b) kNN regression on UBayFS features

(c) linear regression on RENT features

(d) linear regression on UBayFS features

Fig. 5. Predictive performances (on test set) of models trained after feature selection for different numbers of features.

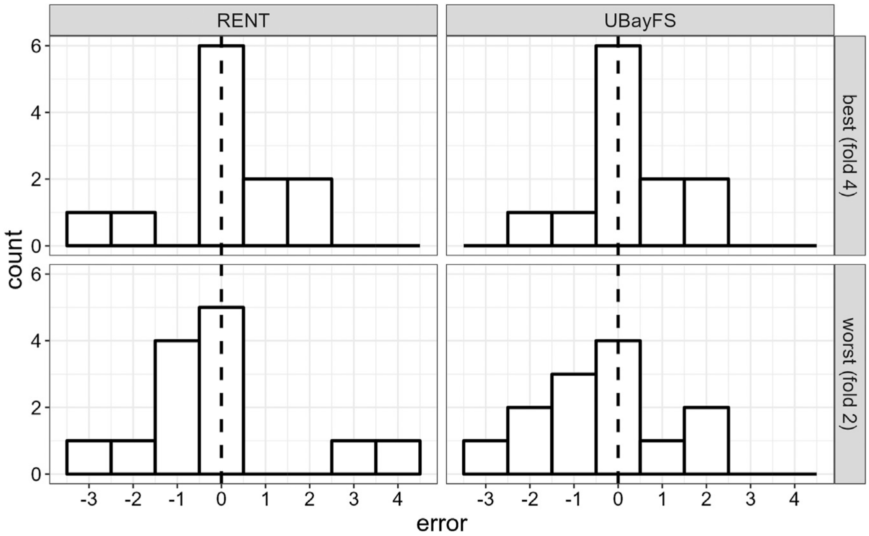


Fig. 6. Histograms of errors on the test set (predicted value by kNN - ground truth) of the folds performing best (fold 4) and worst (fold 2) at $\max_s = 20$ features.

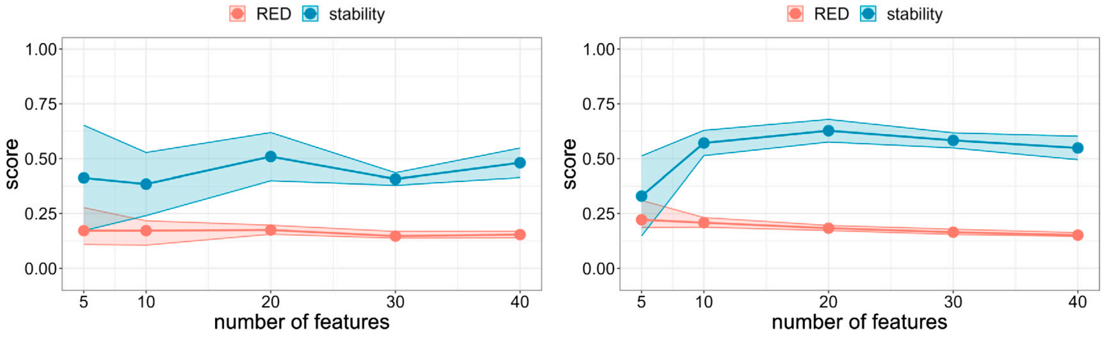
that all methods are able to estimate the intercept correctly. In both folds, the prediction model was able to predict the correct lifetime category for almost half of the patients in the test set. However, the histogram indicates that predictive models based on both feature selectors overestimate the lifetime in test fold 4 (positive errors), while lifetimes in test fold 2 are rather slightly overestimated (negative errors). The main difference in performance between fold 2 and fold 4 is driven by dispersion, i.e. by a minority of patients, which show a high error — due to the small sample size, even a small number of such outliers can impact the total RMSE significantly.

When considering patients with absolute residual values > 2.5 as outliers, RENT shows three outliers in fold 2 (two positive and one negative) and one in fold 4 (negative), while UBayFS shows one outlier in fold 2 (negative). All outliers refer to different patients.

Stability In addition to the performance evaluation, we further investigate qualitative aspects of the selected feature sets, as shown in Fig. 7. The demonstrated stabilities and redundancy rates (RED) of the feature sets selected by RENT and UBayFS across the five folds tend to increase with $\max_s$. While RENT has a slightly lower and more fluctuating stability (around 0.5), UBayFS shows a clear convergence at around 0.6. The RED is below 0.25 for all possible numbers of features, indicating that both RENT and UBayFS select features with small correlations.

3.4. Experiment 2: feature selection with prior knowledge

Previous research on GEP NEN shows that some features impact the survival of patients; those are *Age at diagnosis, WHO performance status, Primary tumor location, Tumor morphology, Tumor differentiation, Lactate*



(a) RENT features (b) UBayFS features

Fig. 7. Stabilities and redundancy rates (RED) of feature sets selected by RENT and UBayFS ($\max_s = 20$ features, each).

dehydrogenase (LDH), Platelets, Albumin, Ki-67, $SUV_{\max}$, and *TNM-staging* [7,10–16]. Tumor differentiation is highly correlated to tumor morphology, so we do not include the feature in this work. Furthermore, findings by [10] indicate a high relevance of the features *Total MTV [cm^3] and Total TLG [g]*, which shall be investigated.

In this experiment, we focus on these features (a total number of 22 features in the encoded space) within our feature selection and prediction pipeline. In particular, during experiment 1, the aforementioned features comprise PERC= 30% of the final feature sets (on average across the five folds and given $\max_s = 20$ features, each). We refer to this score as PERC (percentage of selected features supported by literature). In the following, we deploy prior weights on these features to investigate how UBayFS as a hybrid feature selector combining information from experts and data, performs in comparison to the pure data-driven methods presented in experiment 1. Since RENT cannot incorporate prior feature importances, this evaluation is restricted to UBayFS.

Specifically, we increase the prior weight of the 22 features supported by literature (referred to as prior-elevated features) to the following levels: $w \in \{0.1, 10, 20, \dots, 100, 110\}$ — after evaluating all levels with respect to predictive performance, we restrict to special cases $w = 0.1$ (non-informative prior weighting), $w = 50$ (mediocre prior weighting), and $w = 110$ (strong prior weighting). After applying UBayFS with the given levels of prior information, we examine how the feature set and the predictive performance develop. The case of 0.1 is equivalent to the uniform case without prior knowledge (default setup for UBayFS in experiment 1). In contrast, prior weight 110 indicates that each prior-elevated feature already is assigned a higher score than the maximum score that can be achieved throughout the elementary models ($M = 100$) — as a result, the selected features are exclusively restricted to those with prior information and elementary feature selectors in UBayFS are only used to select a feature set of $\max_s = 20$ features among the 22 prior-elevated features.

Predictive performance Fig. 8 shows the average performances along with the standard deviations across the 5 test folds. In general, lower levels of prior weights do not significantly impact the performance. By increasing the prior weight to a higher level, performance levels lead to stronger variability and an increase of RMSE in the better-performing folds, such as fold 4. Finally, if the prior weight is set to the maximum level of 110, all folds converge to a similar level since the data-driven feature selection hardly contributes to these setups. Thus, a potential conclusion is that moderate levels of prior knowledge can slightly increase models' capabilities. In contrast, strong prior knowledge leads to a convergence towards the global mean performance across all folds — such prior setup acts as a strong restriction of the search space exploited by the feature selector.

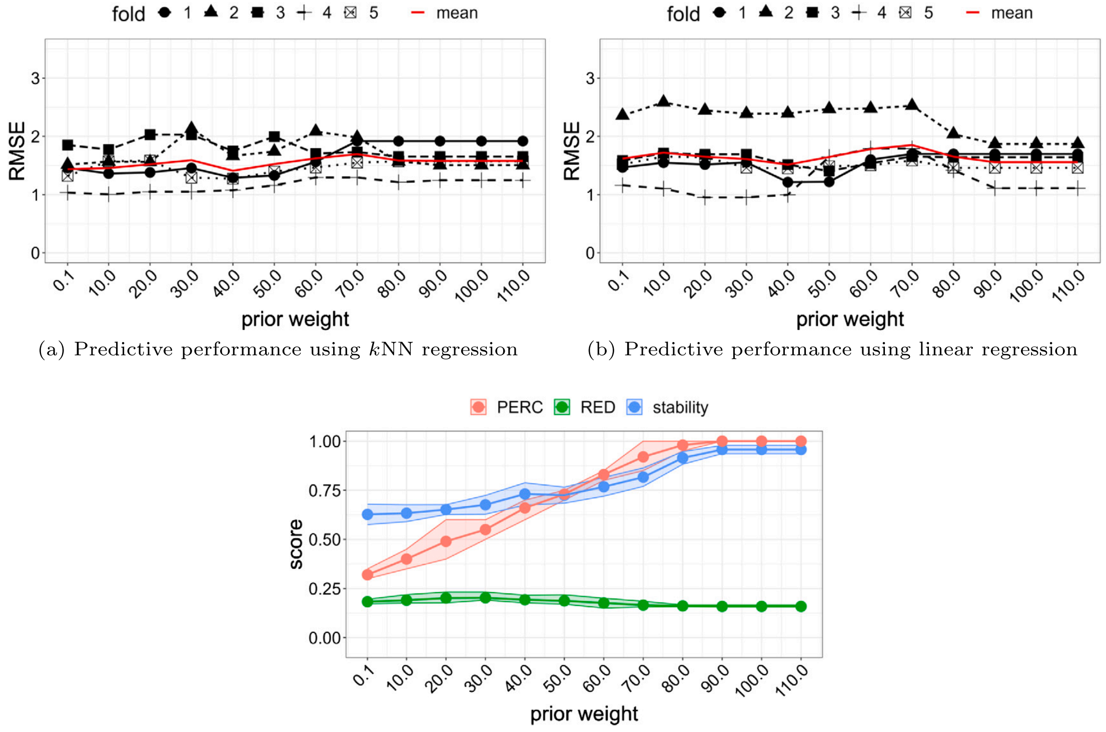
Stability In contrast to the minor effects of prior knowledge on predictive performance, stability increases significantly, as shown in Fig. 8. Finally, at a maximum level of $w = 110$, stability converges towards an almost perfectly stable solution. This is due to the restriction of the search space to the prior-elevated features, which results in a selection of 20 out of only 22 features in total. As expected, the percentage of selected features supported by literature (PERC) also increases linearly with the level of prior weights provided. The redundancy rate between the selected features shows a slight decrease, indicating that the prior-elevated features contain only small correlations.

4. Discussion

Benefits and drawbacks of RENT and UBayFS Both feature selectors evaluated in this work, RENT, and UBayFS, have been designed to satisfy the requirements of medical datasets. Both are ensemble models, which allows them to obtain stable solutions. However, a direct comparison of the two methods from a statistics perspective yields notable differences: Besides the obvious advantage that UBayFS is capable of integrating expert knowledge, RENT restricts the elementary model type to generalized linear models with regularization, while UBayFS is of a generic type, i.e. can be used with an arbitrary elementary feature selection type. As a result, UBayFS is not bounded to a linear model structure but has been deployed with mRMR [34] in this study, which increases the generality of the model. On the other hand, RENT is known to have advantages over UBayFS in purely data-driven setups since the selection criteria take more internal information from the elementary models into account (distribution of parameter coefficients) rather than a pure selection frequency. Thus, in general, RENT has been observed to achieve a higher predictive performance in multiple setups [27] while UBayFS offers more flexibility.

Experiment 1 In our first experiment, we left out prior expert knowledge and let the feature selection be purely data-driven. We know that certain features were prognostic for survival in earlier studies, as mentioned in experiment 2 below. We wanted to study whether the same prognostic features would still be selected and if there were any currently unknown prognostic features that could be further researched. Comparing the two first columns in Table 4 we can see which features are selected repeatedly in different folds with RENT and UBayFS. We must keep in mind that we cannot directly compare the importance of the features in terms of a coefficient (e.g. similar to Cox regression), just that they are repeatedly selected in each fold.

Further, the correlation between features must also be considered when comparing the importance of features which we can find in Fig. 2: the presence of highly correlated features makes the feature selection problem ill-posed, i.e. two distinct sets of features may lead to almost the same result in terms of predictive performance. Intuitively, this issue is caused by the fact that two or more features with a high correlation contribute the same information to the model, and thus, models may choose one over the other by chance in such cases. This results in a lower selection frequency for both features, and thus in an underrat-



(a) Predictive performance using kNN regression

(b) Predictive performance using linear regression

(c) Qualitative evaluation

Fig. 8. Experiment 2: predictive performances on fold 1-5, and qualitative metrics of features sets produced by UBayFS at different levels of prior knowledge on features with evidence from literature ($\max_s = 20$).

ing of their importances to the model, which is why groups of highly correlated features should be analyzed in common.

In block (p) (baseline patient characteristics), we have a few features that one would expect to be prognostic for OS. One obvious one is *Stage IV* disease which does not seem to be chosen at all by RENT and UBayFS. However, looking at the correlation heatmap in Fig. 2 we see that this feature is highly correlated to several other features, among those *Metastatic Disease at Time of Diagnosis* and *Treatment Intention Palliative*. We see that this last one gets chosen four out of five times with both, RENT and UBayFS, which probably explains why *Stage IV* does not seem to be important. Having a palliative treatment intention usually means you have stage IV disease. This is also a well-known prognostic indicator from the literature [8]. Bone metastasis is usually a poor prognostic indicator in several types of cancers [67], and it is not surprising that this is chosen all the time. We also know that *WHO PS* is a prognostic indicator in these patients. This is also reflected in the number of folds it is chosen by RENT and UBayFS, but it is only WHO level 2 that seems to be important. That said, Fig. 2 shows that WHO levels 3 and 4 are highly correlated to some of the SUV parameters which might contribute to those never being selected. *Radical Surgery* is quite often chosen by both RENT and UBayFS and is also a predictable prognostic indicator. Having radical surgery means that all viable tumors are removed, and that is only possible if you have a low tumor burden. This underlines the importance of surgery in the curative intended treatment of this type of cancer.

Next, in block (b) (baseline blood values), we see that both *CRP* and *ALP > Normal <= 3UNL* get selected almost equally many times by both RENT and UBayFS, and both have a high number indicating importance over the other features in this block. A high CRP at baseline has previously been shown to be a poor prognostic feature in some studies [7,68,69], whilst others have not replicated this [70]. This is probably not surprising as this has been shown to be a poor prognostic indicator in advanced cancer patients in a palliative setting, and especially in GEP NEN [71–74]. *ALP* has also been shown in studies to be prognostic for a shorter OS [7,75,76]. For *Albumin* and *Platelets*, RENT chose these only half as many times as UBayFS. Both have been shown to be prognostic indicators of OS [7,8]. Interestingly, *Haemoglobin*, *WBC*, *LDH*, and *Chromogranin A* are barely chosen or are chosen neither by RENT nor UBayFS. All these features have previously been shown to be prognostic for OS [7].

Moving on to block (h) (re-evaluated histology), we have quite a few features that are well-known prognostic indicators. The strongest one from the literature is probably *Ki-67* which is used in the classification system of NEN. The second strongest is probably *Tumor Morphology* which has been shown in several studies to be prognostic for OS [7,8]. We see that *Ki-67* is chosen every time from all five folds both for RENT and UBayFS supporting this feature as a strong prognostic indicator for OS. Further, *Tumor Morphology* gets chosen four out of five times with RENT and three out of five times with UBayFS. This is also to be expected since we know that patients with NET G3 have a better OS than those patients with NEC [77]. What is surprising is that most tumor sites, especially those patients with unknown primary and esophagus NEN, are not chosen by RENT or UBayFS. *Primary Tumor Site* has been shown to be prognostic in several studies [7,8]. Several of the features like *Stroma, Architecture, Vessel Pattern, Co-existing neoplasm*, and *Geographic Necrosis* are considered typical for either NET G3 or NEC [78], and one might assume these are highly correlated with *Tumor Morphology*. Although this is not reflected in Fig. 2. Almost none of these features are chosen with RENT or UBayFS except for *Stroma*. NET G3 typically have hyalinized stroma, and NEC have desmoplastic stroma [78].

Further, in block (i) (PET/CT imaging) the interesting features are *Total MTV*, *Total TLG*, and the *SUV* parameters. From Fig. 2 and previous literature [10] we know that these features are often (if not always) highly correlated. Hence, the selection of $SUV_{\max}$ *(total)* instead of the other features is probably related to this. Moreover, we know from previous studies [10,13–15] that global measures such as *Total MTV* and *Total TLG* are poor prognostic indicators for OS in these tumors, i.e., associated with a poor prognosis of the patient, but we lack stronger evidence in form of larger studies. Here we see that $SUV_{\max}$ *(total)* is chosen in all five folds both for RENT and UBayFS supporting the previous findings that PET-parameters are good prognostic features of OS.

Finally, in block (t) (treatment), we can see that a few features are selected often. *Chemotherapy treatment with cisplatin/etoposide* is not surprisingly a predictor for OS, and most of the patients did indeed receive this combination. No chemotherapy is obviously detrimental. We also see that the *Chemotherapy treatment with temozolomide/everolimus* gets chosen often both by RENT and UBayFS. This is probably because this chemotherapy regimen is more often chosen for those patients with a low *Ki-67*, and these are more likely to be NET G3 which already have a better OS. Further, both *Number of Courses* and *Progression* are two features that are selected often by RENT and UBayFS. *Progression* and *No Progression* are obviously poor prognostic indicators, and one could assume that the higher *Number of Courses* a patient receives, the longer before they have progression, and hence they live longer. This is, of course, only an assumption and interpretation of the data at hand. It is a bit surprising that the response evaluation results did not get chosen. One would assume that patients with the best response - stable disease would fare better than those with progressive disease. Looking at Fig. 2, the features from this block have low correlation coefficients.

Experiment 2 Here we added prior expert knowledge and assigned two different weights. A weight $w = 50$ means approximately 50% expert-driven and 50% data-driven. A weight $w = 110$ means almost purely an expert-driven approach where we effectively force the selection of features only from the subset of those from prior expert knowledge. We concentrated on features that are well documented in several previous studies, although there exist more features in the literature suggesting prognostic values than these. The features selected from prior expert knowledge are listed in the first paragraph in Section 3.4 and marked by an asterisk in Table 4.

If we concentrate on the second, third, and fourth columns, which show the difference between a data-driven, a hybrid, and an expert-driven setup, respectively, we see that none of the marked features drops in importance as we increase the level of expert knowledge. Some features that were never chosen with a pure data-driven model are still not chosen. One could argue that these are probably not strong features to begin with, or that other features contain the same and/or stronger information. A few features only get chosen when almost completely removing the data-driven part and make a huge leap from not being chosen to being chosen five times. We argue that one should be careful to draw conclusions from these features being selected at $w = 110$, as these are forced to be chosen based on expert intervention.

A few features stand out by being stable across all values of w; *WHO Performance Status*, *Albumin*, *Platelets*, *Ki-67*, *Tumor Morphology*, *Total MTV*, *Total TLG*, and $SUV_{\max}$. It would be bold to assume that these features are the most important and stable predictors of OS from the subset of expert knowledge markers, but this is only conditioned on the particular choice of methods presented in this work and requires confirmation in a larger-scale analysis. Further, it is also interesting to notice that even though several parameters from PET are highly correlated, several are still chosen very often by the model. This is in line with the results of our previous study [10]. Moreover, it is a bit surprising that *Primary Tumor Site*, especially *Unknown Primary* and *Esophagus*, is not chosen more often as these are well-known negative predictors of OS [7,8], unless a high level of expert knowledge is applied.

We also notice that some of the other non-marked features drop in importance as we increase w, and this is probably related to the fact that the features overlap in the information they add to the model. A few of these features are also moderately or highly correlated. E.g. *CRP* is correlated with quite a few of the other blood markers, and this could explain why it falls in importance when increasing w. *Mets Bone* (bone metastases) is not listed in the correlation heatmap and thus has no moderate or high correlations with other features, but still completely falls out. Bone metastases usually occur late in several cancers and is a poor prognostic feature. Hence, one should assume that this feature and similar ones like *CRP*, *ALP*, which performs well with low values of w falls of in the pure knowledge-driven model because the model is "forced" to select only marked features. We must remember that the $w = 110$ is an extreme expert-driven model which is probably not clinically relevant but was added to explore and evaluate what the model did in this extreme situation.

In closing, this is a small but novel study with a limited number of patients, and to our knowledge the first study exploring and evaluating RENT and UBayFS on a clinical dataset. Although we cannot ascertain how important different features are compared to each other and if they contribute to poorer or better survival, we find similar results as several previous studies. Both ensemble feature selectors should be evaluated using larger and different patient cohorts, and the evaluation of which w is optimal when using UBayFS should be explored in future studies. The clinical application of both RENT and UBayFS may be validated through already established features from different studies found in the literature, to quickly and effectively compare different features with regard to their prognostic relevance, and to develop or discover novel prognostic features. Clinicians could use this information to deliver patient-tailored survival predictions based on a set of features instead of the traditional median overall survival. Specifically, features extracted from diagnostic imaging can contain the same prognostic information as a traditional pathophysiological feature. Such imaging features are easier to collect as they are non-invasive and all patients go through an imaging work-up before their final diagnosis. Finally, imaging features are less prone to intra- and intertumoral heterogeneity as can be found in several cancers, especially in NEN.

5. Conclusion

From a data science perspective, this work demonstrated the applicability of modern ensemble feature selection techniques RENT and UBayFS for OS prediction on a real-world multi-source dataset from patients diagnosed with high-grade GEP NEN. Overall, our experiments showed that both RENT and UBayFS were able to reduce the number of selected features to only around 18% of the original features (around 20 out of 110 features), while preserving the information needed to achieve a similar predictive performance as in the full dataset with different types of predictive models. In direct comparison, we could conclude that the purely data-based and the hybrid feature selection approach with expert knowledge as input performed equally well in terms of predictive model metrics, while the inclusion of expert knowledge led to a continuous improvement in terms of feature selection stability. The higher the weight on the expert knowledge, the higher the stability, ultimately resulting in an almost purely deterministic selection procedure. This fact underlines that a reasonable trade-off between data and expert knowledge can improve the properties of feature selection results in practical setups. In general, both applied feature selectors allow new insights into a small subset of features, which are highly informative for the overall survival of cancer patients, which underlines the high value of such feature selection methodologies for interpretation in addition to preventing computational issues such as the curse of dimensionality.

From a clinical perspective, we demonstrated the capabilities of modern ensemble feature selectors like RENT and UBayFS for healthcare problems — in particular, the inclusion and comparison of expert- and data-driven setups, as well as combinations of both, allow the user to gain relevant information for clinical use. The most stable and predictive features in our study are *WHO Performance Status*, *Albumin*, *Platelets*, *Ki-67*, *Tumor Morphology*, *Total MTV*, *Total TLG*, and $SUV_{\max}$. This result validates already established known prognostic features and adds support for PET features in prediction for overall survival.

Declaration of competing interest

The authors declare no conflicts of interest.

Acknowledgements

Statement of Ethics. This study was done in concordance with the Declaration of Helsinki. Approval from the regional committee for medical and health research ethics (2012/490, 2012/940, 2018/1940) and the local data protection officer was obtained. Informed consent was obtained from all patients at the time of inclusion but was waived for the patients in the terminal phase and the deceased.

Statement of Funding. This research was in-part supported by the South-Eastern Norway Regional Health Authority (project no. 2018022). It did not receive any specific grant from funding agencies in the commercial or not-for-profit sectors.

References

[1] S. Huang, J. Yang, S. Fong, Q. Zhao, Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges, Cancer Lett. 471 (2020) 61–71, https://doi.org/10.1016/j.canlet.2019.12.007.

[2] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17, https://doi.org/10.1016/j.csbj.2014.11.005.

[3] B.E. White, B. Rous, K. Chandrakumaran, K. Wong, C. Bouvier, M.V. Hemelrijck, G. George, B. Russell, R. Srirajaskanthan, J.K. Ramage, Incidence and survival of neuroendocrine neoplasia in England 1995–2018: a retrospective, population-based study, Lancet Reg. Health, Eur. 23 (2022) 100510, https://doi.org/10.1016/j.lanepe.2022.100510.

[4] R.B. Cetinkaya, B. Aagnes, E. Thiis-Evensen, S. Tretli, D.S. Bergestuen, S. Hansen, Trends in incidence of neuroendocrine neoplasms in Norway: a report of 16, 075 cases from 1993 through 2010, Neuroendocrinology 104 (1) (2015) 1–10, https://doi.org/10.1159/000442207.

[5] International Agency for Research on Cancer, WHO Classification of Tumours. Digestive System Tumours, 1st edition, World Health Organization Classification of Tumours, IARC, 2019.

[6] G. Rindi, D.S. Klimstra, B. Abedi-Ardekani, S.L. Asa, F.T. Bosman, E. Brambilla, K.J. Busam, R.R. de Krijger, M. Dietel, A.K. El-Naggar, L. Fernandez-Cuesta, G. Klöppel, W. McCluggage, H. Moch, H. Ohgaki, E.A. Rakha, N.S. Reed, B.A. Rous, H. Sasano, A. Scarpa, J.-Y. Scoazec, W.D. Travis, G. Tallini, J. Trouillas, J. van Krieken, I.A. Cree, A common classification framework for neuroendocrine neoplasms: an International Agency for Research on Cancer (IARC) and World Health Organization (WHO) expert consensus proposal, Mod. Pathol. 31 (12) (2018) 1770–1786, https://doi.org/10.1038/s41379-018-0110-y.

[7] H. Sorbye, S. Welin, S. Langer, L. Vestermark, N. Holt, P. Osterlund, S. Dueland, E. Hofsli, M. Guren, K. Ohrling, E. Birkemeyer, E. Thiis-Evensen, M. Biagini, H. Gronbaek, L. Soveri, I. Olsen, B. Federspiel, J. Assmus, E. Janson, U. Knigge, Predictive and prognostic factors for treatment and survival in 305 patients with advanced gastrointestinal neuroendocrine carcinoma (WHO G3): the NORDIC NEC study, Ann. Oncol. 24 (1) (2013) 152–160, https://doi.org/10.1093/annonc/mds276.

[8] A. Dasari, K. Mehta, L.A. Byers, H. Sorbye, J.C. Yao, Comparative study of lung and extrapulmonary poorly differentiated neuroendocrine carcinomas: a SEER database analysis of 162, 983 cases, Cancer 124 (4) (2017) 807–815, https://doi.org/10.1002/cncr.31124.

[9] A. Dasari, C. Shen, A. Devabhaktuni, R. Nighot, H. Sorbye, Survival according to primary tumor location, stage, and treatment patterns in locoregional gastroenteropancreatic high-grade neuroendocrine carcinomas, The Oncologist 27 (4) (2022) 299–306, https://doi.org/10.1093/oncolo/oyab039.

[10] H. Langen Stokmo, M. Aly, I.M. Bowitz Lothe, A.J. Borja, S. Mehdizadeh Seraj, R. Ghorpade, X. Miao, G.O. Hjortland, E. Malinen, H. Sorbye, T.J. Werner, A. Alavi, M.-E. Revheim, Volumetric parameters from [18F]FDG PET/CT predicts survival in patients with high-grade gastroenteropancreatic neuroendocrine neoplasms, J. Neuroendocrinol. 34 (7) (2022) e13170, https://doi.org/10.1111/jne.13170, https://onlinelibrary.wiley.com/doi/pdf/10.1111/jne.13170, https://onlinelibrary.wiley.com/doi/abs/10.1111/jne.13170.

[11] M. Heetfeld, C.N. Chougnet, I.H. Olsen, A. Rinke, I. Borbath, G. Crespo, J. Barriuso, M. Pavel, D. O'Toole, T. Walter, Other knowledge network members, characteristics and treatment of patients with G3 gastroenteropancreatic neuroendocrine neoplasms, Endocr.-Relat. Cancer 22 (4) (2015) 657–664, https://doi.org/10.1530/erc-15-0119.

[12] S. Han, H.S. Lee, S. Woo, T.-H. Kim, C. Yoo, B.-Y. Ryoo, J.-S. Ryu, Prognostic value of 18F-FDG PET in neuroendocrine neoplasm, Clinical Nuclear Medicine Publish Ahead of Print, https://doi.org/10.1097/rlu.0000000000003682.

[13] D.L. Chan, E.J. Bernard, G. Schembri, P.J. Roach, M. Johnson, N. Pavlakis, S. Clarke, D.L. Bailey, High metabolic tumour volume on 18-fluorodeoxyglucose positron emission tomography predicts poor survival from neuroendocrine neoplasms, Neuroendocrinology 110 (11–12) (2019) 950–958, https://doi.org/10.1159/000504673.

[14] H.S. Kim, J.Y. Choi, D.W. Choi, H.Y. Lim, J.H. Lee, S.P. Hong, Y.S. Cho, K.-H. Lee, B.-T. Kim, Prognostic value of volume-based metabolic parameters measured by 18F-FDG PET/CT of pancreatic neuroendocrine tumors, Eur. J. Nucl. Med. Mol. Imaging 48 (3) (2014) 180–186, https://doi.org/10.1007/s13139-013-0262-0.

[15] S.M. Lim, H. Kim, B. Kang, H.S. Kim, S.Y. Rha, S.H. Noh, W.J. Hyung, J.-H. Cheong, H.-I. Kim, H.C. Chung, M. Yun, A. Cho, M. Jung, Prognostic value of 18F-fluorodeoxyglucose positron emission tomography in patients with gastric neuroendocrine carcinoma and mixed adenoneuroendocrine carcinoma, Ann. Nucl. Med. 30 (4) (2016) 279–286, https://doi.org/10.1007/s12149-016-1059-x.

[16] G. Centonze, P. Maisonneuve, N. Prinzi, S. Pusceddu, L. Albarello, E. Pisa, M. Barberis, A. Vanoli, P. Spaggiari, P. Bossi, L. Cattaneo, G. Sabella, E. Solcia, S.L. Rosa, F. Grillo, G. Tagliabue, A. Scarpa, M. Papotti, M. Volante, A. Mangogna, A.D. Gobbo, S. Ferrero, L. Rolli, E. Roca, L. Bercich, M. Benvenuti, L. Messerini, F. Inzani, G. Pruneri, A. Busico, F. Perrone, E. Tamborini, A. Pellegrinelli, K. Kankava, A. Berruti, U. Pastorino, N. Fazio, F. Sessa, C. Capella, G. Rindi, M. Milione, Prognostic factors across poorly differentiated neuroendocrine neoplasms: a pooled analysis, Neuroendocrinology, https://doi.org/10.1159/000528186.

[17] S. Ghosh, S. Maulik, S. Chatterjee, I. Mallick, N. Chakravorty, J. Mukherjee, Prediction of survival outcome based on clinical features and pretreatment 18fdg-PET/CT for HNSCC patients, Comput. Methods Programs Biomed. 195 (2020) 105669, https://doi.org/10.1016/j.cmpb.2020.105669.

[18] M.D. Ganggayah, N.A. Taib, Y.C. Har, P. Lio, S.K. Dhillon, Predicting factors for survival of breast cancer patients using machine learning techniques, BMC Med. Inform. Decis. Mak. 19 (1) (2019), https://doi.org/10.1186/s12911-019-0801-4.

[19] S. Mirniaharikandehei, M. Heidari, G. Danala, S. Lakshmivarahan, B. Zheng, Applying a random projection algorithm to optimize machine learning model for predicting peritoneal metastasis in gastric cancer patients using CT images, Comput. Methods Programs Biomed. 200 (2021) 105937, https://doi.org/10.1016/j.cmpb.2021.105937.

[20] L. Brunese, F. Mercaldo, A. Reginelli, A. Santone, An ensemble learning approach for brain cancer detection exploiting radiomic features, Comput. Methods Programs Biomed. 185 (2020) 105134, https://doi.org/10.1016/j.cmpb.2019.105134.

[21] P. Kubben, M. Dumontier, A. Dekker, Fundamentals of Clinical Data Science, Springer International Publishing, 2019.

[22] M.L. Welch, C. McIntosh, B. Haibe-Kains, M.F. Milosevic, L. Wee, A. Dekker, S.H. Huang, T.G. Purdie, B. O'Sullivan, H.J. Aerts, D.A. Jaffray, Vulnerabilities of radiomic signature development: the need for safeguards, Radiother. Oncol. 130 (2019) 2–9, https://doi.org/10.1016/j.radonc.2018.10.027.

[23] D. Wallis, I. Buvat, Clever Hans effect found in a widely used brain tumour MRI dataset, Med. Image Anal. 77 (2022) 102368, https://doi.org/10.1016/j.media.2022.102368.

[24] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A, Math. Phys. Eng. Sci. 374 (2065) (2016) 20150202, https://doi.org/10.1098/rsta.2015.0202.

[25] N. Cueto-López, M.T. García-Ordás, V. Dávila-Batista, V. Moreno, N. Aragonés, R. Alaiz-Rodríguez, A comparative study on feature selection for a risk prediction model for colorectal cancer, Comput. Methods Programs Biomed. 177 (2019) 219–229, https://doi.org/10.1016/j.cmpb.2019.06.001.

[26] T. Emura, S. Matsui, H.-Y. Chen, compound.Cox: univariate feature selection and compound covariate for predicting survival, Comput. Methods Programs Biomed. 168 (2019) 21–37, https://doi.org/10.1016/j.cmpb.2018.10.020.

[27] A. Jenul, S. Schrunner, J. Pilz, O. Tomic, A user-guided Bayesian framework for ensemble feature selection in life science applications (UBayFS), Mach. Learn. 111 (10) (2022) 3897–3923.

[28] C.E. Charlton, M.T. Poon, P.M. Brennan, J.D. Fleuriot, Development of prediction models for one-year brain tumour survival using machine learning: a comparison of accuracy and interpretability, Comput. Methods Programs Biomed. 233 (2023) 107482, https://doi.org/10.1016/j.cmpb.2023.107482.

[29] S. Pozzoli, A. Soliman, L. Bahri, R.M. Branca, S. Girdzijauskas, M. Brambilla, Domain expertise–agnostic feature selection for the analysis of breast cancer data, Artif. Intell. Med. 108 (2020) 101928.

[30] B. Remeseiro, V. Bolon-Canedo, A review of feature selection methods in medical applications, Comput. Biol. Med. 112 (2019) 103375, https://doi.org/10.1016/j.compbiomed.2019.103375.

[31] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc., Ser. B, Methodol. 58 (1) (1996) 267–288.

[32] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, Classification and Regression Trees, Taylor & Francis, 1984.

[33] X. He, D. Cai, P. Niyogi, Laplacian score for feature selection, in: Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'05, MIT Press, Cambridge, MA, USA, 2005, pp. 507–514.

[34] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, J. Bioinform. Comput. Biol. 3 (02) (2005) 185–205.

[35] S. Nogueira, K. Sechidis, G. Brown, On the stability of feature selection algorithms, J. Mach. Learn. Res. 18 (174) (2018) 1–54, http://jmlr.org/papers/v18/17-514.html.

[36] V. Bolón-Canedo, A. Alonso-Betanzos, Recent Advances in Ensembles for Feature Selection, 1st edition, Intelligent Systems Reference Library, Springer International Publishing, Basel, Switzerland, 2018.

[37] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[38] A. Jenul, S. Schrunner, K.H. Liland, U.G. Indahl, C.M. Futsæther, O. Tomic, RENT—repeated elastic net technique for feature selection, IEEE Access 9 (2021) 152333–152346, https://doi.org/10.1109/ACCESS.2021.3126429.

[39] J.-O. Jung, N. Crnovrsanin, N.M. Wirsik, H. Nienhüser, L. Peters, F. Popp, A. Schulze, M. Wagner, B.P. Müller-Stich, M.W. Büchler, T. Schmidt, Machine learning for optimized individual survival prediction in resectable upper gastrointestinal cancer, J. Cancer Res. Clin. Oncol. 149 (5) (2022) 1691–1702, https://doi.org/10.1007/s00432-022-04063-5.

[40] I. Drozdov, M. Kidd, B. Nadler, R.L. Camp, S.M. Mane, O. Hauso, B.I. Gustafsson, I.M. Modlin, Predicting neuroendocrine tumor (carcinoid) neoplasia using gene expression profiling and supervised machine learning, Cancer 115 (8) (2009) 1638–1650, https://doi.org/10.1002/cncr.24180.

[41] W. Liang, P. Yang, R. Huang, L. Xu, J. Wang, W. Liu, L. Zhang, D. Wan, Q. Huang, Y. Lu, Y. Kuang, T. Niu, A combined nomogram model to preoperatively predict histologic grade in pancreatic neuroendocrine tumors, Clin. Cancer Res. 25 (2) (2019) 584–594, https://doi.org/10.1158/1078-0432.ccr-18-1305.

[42] Y. Zhou, L. Song, J. Xia, H. Liu, J. Xing, J. Gao, Radiomics model based on contrast-enhanced CT texture features for pretreatment prediction of overall survival in esophageal neuroendocrine carcinoma, Front. Oncol. 13 (2023), https://doi.org/10.3389/fonc.2023.1225180.

[43] H. Mi, C. Petitjean, B. Dubray, P. Vera, S. Ruan, Robust feature selection to predict tumor treatment outcome, Artif. Intell. Med. 64 (3) (2015) 195–204.

[44] International Agency for Research on Cancer, WHO Classification of Tumours of the Digestive System, 4th edition, World Health Organization Classification of Tumours, IARC, 2010.

[45] R. Boellaard, M.J. O'Doherty, W.A. Weber, F.M. Mottaghy, M.N. Lonsdale, S.G. Stroobants, W.J.G. Oyen, J. Kotzerke, O.S. Hoekstra, J. Pruim, P.K. Marsden, K. Tatsch, C.J. Hoekstra, E.P. Visser, B. Arends, F.J. Verzijlbergen, J.M. Zijlstra, E.F.I. Comans, A.A. Lammertsma, A.M. Paans, A.T. Willemsen, T. Beyer, A. Bockisch, C. Schaefer-Prokop, D. Delbeke, R.P. Baum, A. Chiti, B.J. Krause, FDG PET and PET/CT: EANM procedure guidelines for tumour PET imaging: version 1.0, Eur. J. Nucl. Med. Mol. Imaging 37 (1) (2010) 181–200, https://doi.org/10.1007/s00259-009-1297-4.

[46] R. Boellaard, R. Delgado-Bolton, W.J.G. Oyen, F. Giammarile, K. Tatsch, W. Eschner, F.J. Verzijlbergen, S.F. Barrington, L.C. Pike, W.A. Weber, S. Stroobants, D. Delbeke, K.J. Donohoe, S. Holbrook, M.M. Graham, G. Testanera, O.S. Hoekstra, J. Zijlstra, E. Visser, C.J. Hoekstra, J. Pruim, A. Willemsen, B. Arends, J. Kotzerke, A. Bockisch, T. Beyer, A. Chiti, B.J. Krause, FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0, Eur. J. Nucl. Med. Mol. Imaging 42 (2) (2014) 328–354, https://doi.org/10.1007/s00259-014-2961-x.

[47] E. Eisenhauer, P. Therasse, J. Bogaerts, L. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, J. Verweij, New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1), Eur. J. Cancer 45 (2) (2009) 228–247, https://doi.org/10.1016/j.ejca.2008.10.026.

[48] A.B. Mariotto, A.-M. Noone, N. Howlader, H. Cho, G.E. Keel, J. Garshell, S. Woloshin, L.M. Schwartz, Cancer survival: an overview of measures, uses, and interpretation, JNCI Monogr. 2014 (49) (2014) 145–186, https://doi.org/10.1093/jncimonographs/lgu024.

[49] J.M. Bland, D.G. Altman, Statistics notes: survival probabilities (the Kaplan-Meier method), BMJ 317 (7172) (1998) 1572–1580, https://doi.org/10.1136/bmj.317.7172.1572.

[50] D.A. Bennett, How can I deal with missing data in my study?, Aust. N. Z. J. Public Health 25 (5) (2001) 464–469, https://doi.org/10.1111/j.1467-842X.2001.tb00294.x, https://www.sciencedirect.com/science/article/pii/S1326020023036488.

[51] M. Kuhn, K. Johnson, Applied Predictive Modeling, 1st edition, Springer, New York, NY, 2013.

[52] R.K. Bania, A. Halder, R-hefs: rough set based heterogeneous ensemble feature selection method for medical data classification, Artif. Intell. Med. 114 (2021) 102049.

[53] S. Pölsterl, S. Conjeti, N. Navab, A. Katouzian, Survival analysis for high-dimensional, heterogeneous medical data: exploring feature extraction as an alternative to feature selection, Artif. Intell. Med. 72 (2016) 1–11.

[54] A. Zheng, Feature Engineering for Machine Learning, O'Reilly Media, Sebastopol, CA, 2018.

[55] I.-K. Yeo, A new family of power transformations to improve normality or symmetry, Biometrika 87 (4) (2000) 954–959, https://doi.org/10.1093/biomet/87.4.954.

[56] B. Srujana, D. Verma, S. Naqvi, Machine learning vs. survival analysis models: a study on right censored heart failure data, Commun. Stat., Simul. Comput. (2022) 1–18.

[57] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc., Ser. B, Stat. Methodol. 67 (2) (2005) 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[58] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, 2009.

[59] R Core Team R, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2022, https://www.R-project.org/.

[60] G. Van Rossum, F.L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[61] A. Jenul, S. Schrunner, B.N. Huynh, O. Tomic, RENT: a Python package for repeated elastic net feature selection, J. Open Sour. Softw. 6 (63) (2021) 3323, https://doi.org/10.21105/joss.03323.

[62] A. Jenul, S. Schrunner, UBayFS: an R package for user guided feature selection, J. Open Sour. Softw. 8 (81) (2023) 4848, https://doi.org/10.21105/joss.04848.

[63] M. Kuhn, Caret: classification and regression training, R package version 6.0-93, https://CRAN.R-project.org/package=caret, 2022.

[64] 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, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[65] H. Wickham, Ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag, New York, 2016, https://ggplot2.tidyverse.org.

[66] J. Krawczuk, T. Łukaszuk, The feature selection bias problem in relation to high-dimensional gene data, Artif. Intell. Med. 66 (2016) 63–71.

[67] G. Chen, Q. Xu, S. Qian, Z. Wang, S. Wang, Survival analysis in gastrointestinal neuroendocrine carcinoma with bone metastasis at diagnosis, Front. Surg. 9 (2022), https://doi.org/10.3389/fsurg.2022.820725.

[68] Ömer Komaç, G. Bengi, Özgül Sağol, M. Akarsu, C-reactive protein may be a prognostic factor for the whole gastroenteropancreatic neuroendocrine tumor group, World J. Gasterointest. Oncol. 11 (2) (2019) 139–152, https://doi.org/10.4251/wjgo.v11.i2.139.

[69] A. Nießen, S. Schimmack, M. Sandini, D. Fliegner, U. Hinz, M. Lewosinska, T. Hackert, M.W. Büchler, O. Strobel, C-reactive protein independently predicts survival in pancreatic neuroendocrine neoplasms, Sci. Rep. 11 (1) (2021), https://doi.org/10.1038/s41598-021-03187-x.

[70] P. Freis, E. Graillot, P. Rousset, V. Hervieu, L. Chardon, C. Lombard-Bohas, T. Walter, Prognostic factors in neuroendocrine carcinoma: biological markers are more useful than histomorphological markers, Sci. Rep. 7 (1) (2017), https://doi.org/10.1038/srep40609.

[71] N. Gebauer, M. Ziehm, J. Gebauer, A. Riecke, S. Meyhöfer, B. Kulemann, N. von Bubnoff, K. Steinestel, A. Bauer, H.M. Witte, The Glasgow prognostic score predicts survival outcomes in neuroendocrine neoplasms of the gastro–entero–pancreatic (GEP-NEN) system, Cancers 14 (21) (2022) 5465, https://doi.org/10.3390/cancers14215465.

[72] K. Amano, I. Maeda, T. Morita, T. Miura, S. Inoue, M. Ikenaga, Y. Matsumoto, M. Baba, R. Sekine, T. Yamaguchi, T. Hirohashi, T. Tajima, R. Tatara, H. Watanabe, H. Otani, C. Takigawa, Y. Matsuda, H. Nagaoka, M. Mori, H. Kinoshita, Clinical implications of c-reactive protein as a prognostic marker in advanced cancer patients in palliative care settings, J. Pain Symptom Manag. 51 (5) (2016) 860–867, https://doi.org/10.1016/j.jpainsymman.2015.11.025.

[73] P.C. Hart, I.M. Rajab, M. Alebraheem, L.A. Potempa, C-reactive protein and cancer—diagnostic and therapeutic insights, Front. Immunol. 11 (2020), https://doi.org/10.3389/fimmu.2020.595835.

[74] S. Shrotriya, D. Walsh, A.S. Nowacki, C. Lorton, A. Aktas, B. Hullihen, N. Benanni-Baiti, K. Hauser, S. Ayvaz, B. Estfan, Serum C-reactive protein is an important and powerful prognostic biomarker in most adult solid tumors, PLoS ONE 13 (8) (2018) e0202555, https://doi.org/10.1371/journal.pone.0202555.

[75] T.E. Clancy, T.P. Sengupta, J. Paulus, F. Ahmed, M.-S. Duh, M.H. Kulke, Alkaline phosphatase predicts survival in patients with metastatic neuroendocrine tumors, Dig. Dis. Sci. 51 (5) (2006) 877–884, https://doi.org/10.1007/s10620-006-9345-4.

[76] M. Ter-Minassian, J.A. Chan, S.M. Hooshmand, L.K. Brais, A. Daskalova, R. Heafield, L. Buchanan, Z.R. Qian, C.S. Fuchs, X. Lin, D.C. Christiani, M.H. Kulke, Clinical presentation, recurrence, and survival in patients with neuroendocrine tumors: results from a prospective institutional database, Endocr.-Relat. Cancer 20 (2) (2013) 187–196, https://doi.org/10.1530/erc-12-0340.

[77] H. Sorbye, E. Baudin, A. Perren, The problem of high-grade gastroenteropancreatic neuroendocrine neoplasms, Endocrinol. Metab. Clin. N. Am. 47 (3) (2018) 683–698, https://doi.org/10.1016/j.ecl.2018.05.001.

[78] H. Elvebakken, A. Perren, J.-Y. Scoazec, L.H. Tang, B. Federspiel, D.S. Klimstra, L.W. Vestermark, A.S. Ali, I. Zlobec, T.Å. Myklebust, G.O. Hjortland, S.W. Langer, H. Gronbaek, U. Knigge, E.T. Janson, H. Sorbye, A consensus-developed morphological re-evaluation of 196 high-grade gastroenteropancreatic neuroendocrine neoplasms and its clinical correlations, Neuroendocrinology 111 (9) (2020) 883–894, https://doi.org/10.1159/000511905.