Curvature-based Feature Selection with Application in Classifying Electronic Health Records

Zheming Zuo[*][†]
Department of Computer Science
Durham University
Durham DH1 3LE, UK
zheming.zuo@durham.ac.uk

Jie Li[*]
School of Computing, Engineering & Digital Technologies
Teesside University
Middlesbrough TS3 6DR, UK
jie.li@tees.ac.uk

Han Xu
Institute of Microelectronics, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing 100029, China
ann.gong.qifeng@gmail.com

Noura Al Moubayed
Department of Computer Science
Durham University
Durham DH1 3LE, UK
noura.al-moubayed@durham.ac.uk

ABSTRACT

Disruptive technologies provides unparalleled opportunities to contribute to the identifications of many aspects in pervasive healthcare, from the adoption of the Internet of Things through to Machine Learning (ML) techniques. As a powerful tool, ML has been widely applied in patient-centric healthcare solutions. To further improve the quality of patient care, Electronic Health Records (EHRs) are commonly adopted in healthcare facilities for analysis. It is a crucial task to apply AI and ML to analyse those EHRs for prediction and diagnostics due to their highly unstructured, unbalanced, incomplete, and high-dimensional nature. Dimensionality reduction is a common data preprocessing technique to cope with high-dimensional EHR data, which aims to reduce the number of features of EHR representation while improving the performance of the subsequent data analysis, *e.g.* classification. In this work, an efficient filter-based feature selection method, namely Curvature-based Feature Selection (CFS), is presented. The proposed CFS applied the concept of Menger Curvature to rank the weights of all features in the given data set. The performance of the proposed CFS has been evaluated in four well-known EHR data sets, including Cervical Cancer Risk Factors (CCRFDS), Breast Cancer Coimbra (BCCDS), Breast Tissue (BTDS), and Diabetic Retinopathy Debrecen (DRDDS). The experimental results show that the proposed CFS achieved state-of-the-art performance on the above data sets against conventional PCA and other most recent approaches. The source code of the proposed approach is publicly available at https://github.com/zhemingzuo/CFS.

Keywords feature selection · precision medicine · healthcare · electronic health records · classification

1 Introduction

Disruptive technologies (DTs), including Industry 4.0, Artificial Intelligence (AI), Machine Learning (ML), Big Data, the Internet of Things (IoT), Virtual Reality (VR), etc., are innovations that significantly alters the way that consumers, industries or businesses operate [1]. Among those disruptive technologies, AI and ML are powerful tools to make computers learn and mimic human's thinking, and thus to help analyse the huge amount of medical data to identify potential health issues, such as [2, 3].

Due to the era of big data, large amounts of high-dimensional data have become available in a variety of domains, especially within the realm of digital healthcare [4]. The dramatically increased data volume has become a challenge

---

[*]Equal contribution.
[†]Corresponding author.

for effective and efficient data analysis, as it significantly increases the memory storage requirement and computational costs [5]. To improve the quality of patient care more efficiently, Electronic Health Records (EHRs) are widely employed in the healthcare facilities for analysis. Currently, maintenance of such EHRs has become a crucial task in the medical sector. Patients' digital healthcare data is usually highly unstructured and consists of various features and diagnostics-related information. In addition, EHR data may include missing values and a certain degree of redundancy. Due to the incompleteness, imbalance, inherent heterogeneity, and high-dimensional features of EHRs, it is essential to apply disruptive technologies to provide effective and low-cost data analysis methods, and thus to explore such healthcare data for data mining and data analytics purposes.

Dimensionality reduction is an efficient data preprocessing technique for the analysis of high-dimensional data that aims to reduce the number of features while improving the classification performance (*e.g.* treatment planning [6], survival analysis [7], and risk prediction [8]) and reducing the related computational cost [9]. It is important to identify the most significant factors that related to disease, which helps in removing unnecessary and redundant data from the given data sets, thus increasing the data analysis performance. The selection is usually achieved by either projecting the original data into a lower feature space, *a.k.a.* feature extraction [10], or selecting a subset of features from the original data, *i.e.* feature selection [11]. For the latter, the least relevant features that are required to be removed can be identified by two criteria: a) features that are not correlated with any other features (*i.e.* redundancy); b) features that do not contribute to the classification decision (*i.e.* noise).

It is well-known that the dimensionality reduction-based feature extraction approaches, such as Principal Component Analysis (PCA), reduces the number of features by mapping the original data set into a new feature space with lower dimensions, which changes or removes the physical meanings of original features. For instance, a low-complexity algorithm, which is developed based on Discrete Wavelet Transform (DWT), for electrocardiogram (ECG) data feature extraction, is proposed in [12]. In addition, to deal with the text classification problems, the Latent Semantic Index (LSI) [13] and PCA algorithms have been employed to extract the text feature and have achieved good results [14]. In contrast, methods of selecting a subset of features keep the physical meanings of the original features and enable models with better interpretability, but the underlying complexity of the subset evaluation metric may lead to unnecessary computational cost [15]. For example, a time series feature selection method is proposed to filter the available features with respect to their significance for the regression task [16]. This motivates us to design an efficient selection-based feature selection method that could meet the requirement of the real-time system in the era of big data.

Recently, pervasive healthcare has become a central topic which has attracted intensive attention from academia, industry, as well as healthcare sectors [17, 18, 19, 20, 21, 22, 23, 24]. In this problem domain, highly class-imbalanced data sets with a large number of missing values are common problems [25]. It has been proved that the selected features might have a higher degree of usefulness in comparison with the projected features, due to preservation of the original semantics of the dimensions [9, 11]. Thus, we focus on selecting a sub-set of features, even using the anonymised data set (*e.g.* one possible reason for having the missing attribute values could be that the participants or patients are reluctant to share personally identifiable information with the public [25]), for efficient medical data classification.

Based on the aforementioned two motivations, we address the issues of time complexity and efficiency in an intuitive, explainable fashion in this work. Our contribution is two-fold:

1. A filter-based feature selection method, called Curvature-based Feature Selection (CFS), is proposed to select discriminative attributes in line with the ranked and averaged curvature values for each dimension in the given EHR data set.

2. The presented CFS approach is also embedded into a previously proposed fuzzy inference system, TSK+ [26, 27], to constitute CFS-TSK+, for supporting better decision-making of clinical diagnosis, *i.e.* improving the performance of classifying digital healthcare data.

The rest of the paper is structured as follows. Section 2 expresses the related works. Section 3 proposes our CFS approach and CFS-TSK+ classifier. Section 4 discusses the experimental results, and the paper is concluded in Section 5.

2 Background

In this section, the most recent developments of Machine Learning techniques in classifying medical data will be showcased first. This is followed by revisiting dimensionality reductions techniques for EHR data from the perspectives of feature extraction and feature selection, respectively.

2.1 Machine Learning for Digital Healthcare

In the past few decades, Machine Learning and deep learning algorithms have been widely proposed for solving healthcare problems, such as diagnosis prediction of various diseases including cervical cancer [28], breast cancer [29], and thoracic disease [30], which have usually taken the form of classification.

Due to privacy considerations [31, 32, 33], there is a large number of healthcare data sets containing missing values. To cope with this common issue, the Bayesian Possibilistic C-means (BPCM) [17] was devised to interpolate the missing values by extending the Fuzzy C-Means clustering algorithm (to model the noise and uncertainty) with the support of Bayesian theory (to calculate cluster centroids). The Gene Sequence-based Auxiliary Model (GSAM) [18], as an ensemble learner, was proposed to predict the missing values via data correction and classify testing data samples via a combination of multiple weak learners within a gene auxiliary module.

To enhance the classification performance in terms of accuracy, the Weight Constrained Neural Network (WCNN) was proposed [19]. WCNN utilises network training to solve a constraint optimisation problem. An extension of the Broad Learning System (BLS) was devised by adding a label-based autoencoder (BLS II), to learn robust feature representations in an ensemble way, and also to tune the hyper-parameters in a Support Vector Machine (SVM), namely BLS II-SVM [20]. Another ensemble learner, VIBES, was presented to detect the dependency between attributes in the given data set and to speed up the search for base learners [21].

In addition, Genetic Algorithm has been adopted to optimise the performance of Random Forest (RFGA) [23]. Work towards the enhancement of activation functions in neural networks was also proposed, such as Variable Activation Function (VAF) [24] and Adaptive Takagi-Sugeno-Kang (AdaTSK) [34]. Apart from those adaptive action functions, a proposition of a two-layer mixture of factor analysers with joint factor loading (2L-MJFA) was presented to conduct the dimensionality reduction and classification together [22]. This is done by utilising two mixtures nested with each other, each of which containing several components, where each class of the data sets is represented in a specific mixture of factor analysers (MFA). Such an approach has been proven to be suitable for small-scale data sets, particularly those that contains a smaller number of data instances but includes a larger number of data attributes.

2.2 Dimensionality Reduction for EHRs

EHR data usually has a high-dimensions, thereby containing a large number of input features. It is noteworthy that some of the input features may not be relevant to the problem to be resolved. To effectively deal with such high-dimensional data, a typical solution is to apply specific techniques to reduce the dimensions of the original data set. Fundamentally, the dimensionality reduction techniques are typically divided into two aspects: 1) feature extraction, which combines the original features and creates a new set of feature representation; and 2) feature selection, which selects a subset of the original features [35]. Fig. 1 depicts the major difference between those two types of techniques, and both technologies are described below.
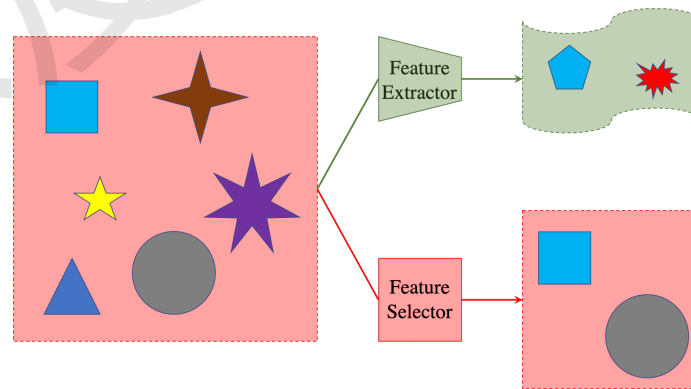

Figure 1: Illustration of feature extraction and feature selection.

2.2.1 Feature Extraction

Feature extraction (FE), also termed as Feature Construction, is a substitute for feature selection that transforms the original data from a high-dimensional space into a low-dimensional one, as illustrated in the upper pathway of Fig. 1. By adopting this type of techniques, the problem is represented in a more discriminating (*i.e.* informative) space, thus leading to a more efficient analysis process. Such techniques have typically been applied in the fields of medical

image analysis, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT) scans, Ultrasound and X-Rays [35]. The common feature extraction techniques can be grouped into two main types: linear and non-linear. Linear feature extraction approaches, such as PCA, adopt the matrix factorisation method to transform the original data into a lower-dimensional subspace. For instance, PCA looks for "principal components" in the given data that are uncorrelated eigenvectors by considering the covariance matrix and its eigenvalues and eigenvectors [36]. Although unsupervised PCA is highly effective in identifying important features of the data, it cannot easily determine the non-linear relationship among the features, which commonly exists in complex EHRs, especially, electrocardiogram (ECG), electroencephalography (EEG) [37], and biological data [35].

Compared with linear feature extraction methods, which linearly map the original data into a low-dimensional subspace, non-linear feature extraction approaches work in different ways to represent the non-linear distribution, *e.g.* Kernel PCA [38], Locally Linear Embedding (LLE) [38], and Self-Organising Maps (SOM) [39]. Such approaches worked based on the hypothesis that the data lies on an embedded non-linear manifold that has a lower dimension than the raw data space that lies within it [38].

Although the extracted features have a higher discriminating power that not only reduces the computational cost but also increases the classification accuracy, the combinations of the newly created sets of attributes may not have real meaning; therefore, those types of approaches may not be good methods with respect to readability, explainability, as well as transparency [35].

2.2.2 Feature Selection

Feature selection (FS) is a process of selecting a subset of the most important/relevant attributes from the given data set to use for model construction (*i.e.* data modelling). Similar to FE, the aim of FS is also to aid in the task of generating accurate predictive models; however, this is achieved by identifying and removing irrelevant and redundant features from the given data set which do not contribute to increasing the performance of a system model or, perhaps, may reduce the accuracy of the predictive model [40], as depicted in the lower pathway of Fig. 1. Therefore, it is perfect when interpretability and knowledge extraction are crucial, *e.g.* in medicine. Essentially, FS methods assess and evaluate the individual features in the original data set to determine the relevance of each feature for the given problem, so as to select the most relevant ones. In general, based on the relationship with the different learning methods, the process of feature selection can be categorised into three types: filter method, wrapper method, and embedded method [41].

- *Filter*: The filter method focuses on the general characteristics of each feature and ranks features based on a certain number of evaluation criteria. This is followed by a threshold value selection process in order to eliminate the features that are less than the selected crisp value. This method is computationally efficient and learning invariant, as it is independent of any learning algorithm. The limitation of such approaches is that there is no interaction between the classifiers, class labels (outputs), and dependency of one feature over others. Consequently, those approaches may fail to determine the most "useful" features [41]. Typical filter feature selection methods include Information Gain (IG) [42], Mutual Information (MI) [43, 44], and Chi-Square Test (CST) [45].

- *Wrapper*: Unlike the filter method, the wrapper method depends on the performance of the learning algorithm to select features. In this method, candidate subsets of features are evaluated by an induction algorithm. The learning algorithms are employed to analyse the relationship between input features and the outputs (*i.e.* class labels), and thus identify the most useful/relevant features. Compared with filter methods, which are not computationally intensive, wrapper approaches usually have a complex progress and are more computationally costly than filter methods. In addition, this method is more prone to over-fitting on small training data sets [41].

- *Embedded*: Though embedded method-based approaches still interact with learning algorithms for selecting relevant features, they conduct a different procedure from the filter and wrapper methods. In general, the embedded approaches can be described as a combination of the filter method and the wrapper method. They not only measure the relations between one input feature and its output feature (*i.e.* class labels) but also considers each feature's general characteristic itself locally for better local discrimination [46]. In particular, the embedded approaches firstly use the independent criteria to determine the optimal feature subsets from the given data set, and then, the learning algorithm is applied to finalise the final optimal feature subsets from the previous results. Compared with the wrapper method, the embedded approaches require low computational cost, and the chance of over-fitting is also reduced [46].

Recently, a hybrid method has also widely been employed to preprocess EHRs, in order to increase the model prediction capability. This method aggregates one or more FS approaches together, for example, filter and wrapper methods combined, to take the advantages of different methods, and hence to generate optimal results. The hybrid method can

usually achieve better performance, *e.g.* higher prediction accuracy; however, it also requires a higher computational cost [47].

3 Proposed System

A novel filter approach feature selection method, called Curvature -based Feature Selection (CFS), is proposed and detailed in this section. The system pipeline is illustrated in Fig. 3, which comprises three main components: two-dimensional (2-D) data re-construction, feature weight calculation by Menger Curvature (depicted in Fig. 2), and feature ranking.

3.1 Menger Curvature

Menger Curvature ($\mathcal{MC}$) [48] measures the curvature of triple data points within the n-dimensional Euclidean space $\mathbb{E}^n$ represented by the reciprocal of the radius of the circle that passes through the three points q_1, q_2, and q_3 in Fig. 2.
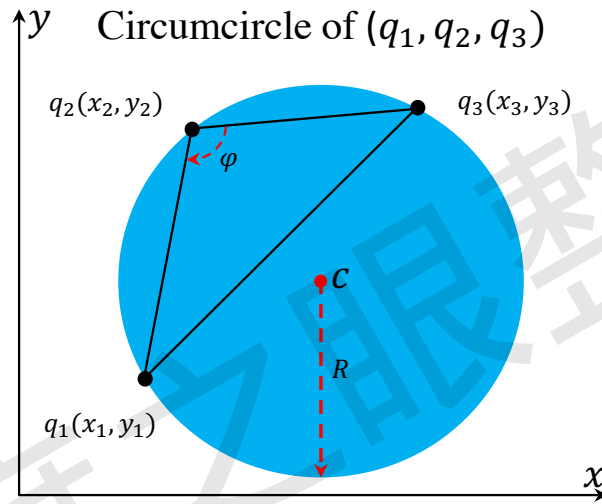

Figure 2: The measurement of the Menger Curvature on a 2-D space.

In this work, only two-dimensional plane curves problems are considered. Given that $q_1(x_1, y_1), q_2(x_2, y_2)$, and $q_3(x_3, y_3)$ are the three points in a 2-D space, $\mathbb{E}^2$ and q_1, q_2, q_3 are non-collinear, as expressed in Fig. 2, and $\mathcal{MC}$ on B is calculated as:

$$\mathcal{MC}(q_1, q_2, q_3) = \frac{1}{R} = \frac{2\sin(\varphi)}{\|q_1, q_3\|}, \tag{1}$$

where R represents the radius, $\|q_1, q_3\|$ denotes the Euclidean distance between q_1 and q_3, and φ is the angle of the q_2-corner of the triangle spanned by q_1, q_2, q_3, which can be calculated in line with the Law of Cosines:

$$\cos(\varphi) = \frac{\|q_1, q_2\|^2 + \|q_2, q_3\|^2 - \|q_1, q_3\|^2}{2 \cdot \|q_1, q_2\|^2 \cdot \|q_2, q_3\|^2}. \tag{2}$$

$\mathcal{MC}$ on points q_1 and q_3 is not calculable, as these points are boundary points. The efficacy of $\mathcal{MC}$ is confirmed in constructing a Mamdani fuzzy rule base [49].

3.2 Curvature-based Feature Selection

Assume that a high-dimensional raw data set, denoted as $\mathcal{U} \in \mathbb{R}^{m \times n}$, contains m data instances, n input attributes, and a single output feature y. In the real-world problem domain, a data cleaning process (*e.g.* removing attributes with
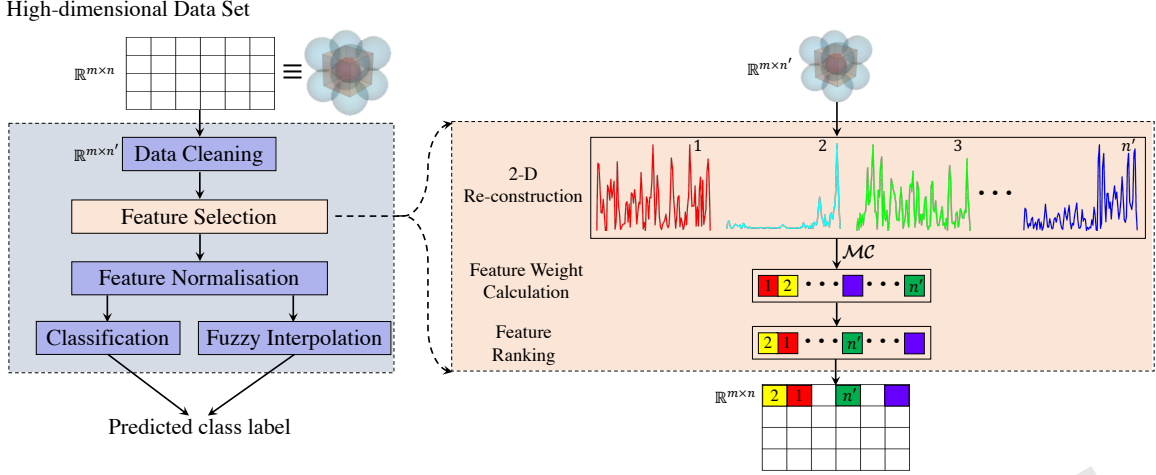
Figure 3: Architecture of the proposed CFS method.

missing values) and a data normalisation phase (*e.g.* bounding all the values within the interval of $[0, 1]$) may be applied on $\mathcal{U}$ to obtain $\mathcal{U}' \in \mathbb{R}^{m \times n'}$ s.t. $n' < n$. In this work, we adopt the Min-Max (MM) normalisation technique:

$$\mathcal{U}' = \frac{\mathcal{U} - \min(\mathcal{U})}{\max(\mathcal{U}) - \min(\mathcal{U})}. \tag{3}$$

This operation helps to cancel out the influence of possible large variations in the raw data set and guarantees that our CFS is able to compare the curvatures for each attribute in an equitable manner. In other words, all the attribute values are normalised to the same frame of reference to ensure the correct rankings are generated by CFS. The proposed CFS method is described as follows:

Step 1 – *2-D Data Re-construction*: The first step of the proposed CFS is to break down the cleaned high-dimensional data set $\mathcal{U}'$ into n' 2-D planes, which is implemented by combining all input attributes, $\mathcal{F}'_i$ ($1 \leqslant i \leqslant n'$), and the output y. Thus, $\mathcal{U}'$ can be decomposed to n' 2-D planes, represented as $\mathcal{P}_{(\mathcal{F}'_i, y)}$.

Step 2 – *Feature Weighting*: For each decomposed 2-D plane, $\mathcal{P}_{(\mathcal{F}'_i, y)}$, the Menger Curvature method, introduced in Section 3.1, is adopted to obtain the averaged curvature value of the feature $\mathcal{F}'_i$. Given that a decomposed 2-D panel ($\mathcal{P}_{(\mathcal{F}'_i, y)}$) contains m data instances, the Menger Curvature value ($\mathcal{MC}^i_{m_j}$) of data point m_j ($2 \leqslant j \leqslant m - 1$) can be determined by Eq. (1). To this end, the mean of $\mathcal{MC}$ for $\mathcal{F}'_i$, denoted as $\widehat{\mathcal{MC}}_{\mathcal{F}'_i}$, is computed as:

$$\widehat{\mathcal{MC}}_{\mathcal{F}'_i} = \frac{1}{m-2} \sum_{j=2}^{m-1} \mathcal{MC}^i_{m_j}, \tag{4}$$

where $\mathcal{MC}^i_{m_j}$ represents the curvature value of the m_j^{th} data point in feature $\mathcal{F}'_i$. $\widehat{\mathcal{MC}}_{\mathcal{F}'_i}$ indicates the corresponding weight of the feature $\mathcal{F}'_i$, the greater value of $\widehat{\mathcal{MC}}_{\mathcal{F}'_i}$ signifies a higher degree of importance of the corresponding feature $\mathcal{F}'_i$ for the data set $\mathcal{U}'$, and vice versa.

It is noted that for a given data set, changing the order of data points leads to a different value of $\mathcal{MC}^i_{m_j}$. However, the order of the all data instances will be changed accordingly. For this reason, the ranking of the features remain the same.

Step 3 – *Feature Ranking and Feature Selection*: A conventional ordinal ranking method is used to rank the features, based on the obtained $\widehat{\mathcal{MC}}_{\mathcal{F}'_i}$. Therefore, the features of $\mathcal{U}'$ are ranked. This is followed by selecting the corresponding features from the raw data set $\mathcal{U}$. Given a threshold ∂, the features with $\widehat{\mathcal{MC}}_{\mathcal{F}'_i}$ greater than the given threshold ∂ will be selected. Equivalently, a $\text{Top}K$ method can be employed:

$$\mathcal{U}'' := \mathcal{U}\left[\text{Rank}^{\text{Top}K}\left(\widehat{\mathcal{MC}}_{\mathcal{F}'_i} \right) \right], \tag{5}$$

such that $\mathcal{U}'' \in \mathbb{R}^{m \times n'}$. To this end, we have reduced the dimensionality of $\mathcal{U}$ to $\mathcal{U}''$ while preserving the statistical nature of the original data set. Then, in line with the rest of the parts shown in Fig. 3, the obtained $\mathcal{U}''$ will be further normalised and classified.

3.3 Feature Normalisation

In this work, we involve two different data normalisation processes: one before curvature calculation, the other after. The first data normalisation process is applied to make sure that the curvatures for each attribute can be compared equitably. The second data normalisation process is optional. It is used to reduce the noise and hence improve the classification performance on EHR data sets.

Concretely, the Min-Max data normalisation technique has been adopted for feature ranking, which aims to ensure each data attribute is compared in an equitable manner. To improve the performance of classification and ensure the degree of membership in the TSK+ is calculable, the selected features in $\mathcal{U}''$ are further normalised using a total number of eight normalisation techniques [34] in this work, including Min-Max (MM) normalisation, ℓ1-normalisation, ℓ2-normalisation, Power Normalisation (PN), and its variants (*i.e.* ℓ1PN, ℓ2PN, PNℓ1, and PNℓ2).

3.4 Classification

To classify the selected and normalised features, nine classifiers [40, 34] are used, namely Gaussian Naïve Bayes (GNB), Random Forest (RF), AdaBoost (AB), Logistic Regression (LR), Linear Support Vector Machine (Linear SVM), Decision Tree (DT), k Nearest Neighbours (kNN), and Back-Propagation Neural Network (BPNN). Additionally, we also combine the proposed CFS method with TSK+ (CFS-TSK+) and evaluate its performance for the classification of four benchmark medical data sets.

4 Experiments

For performance verification and evaluation, the proposed CFS method is compared against PCA, IG, MI, and CST on four benchmark clinical data sets. In the following, we describe the data sets and the experimental setup we used to examine the aforementioned techniques.

4.1 Data sets

Table 1: Data set structures. † denotes the exclusion of class label. * represents the number of dimensions contain the missing values.

Data Set	# of instances	# of dim.†	# of classes	Missing values?	# of dim. *†	Year published
CCRFDS	858	35	2	✔	26	2017
BCCDS	116	9	2	✘	N/A	2018
BTDS	106	9	6	✘	N/A	2010
DRDDS	1,151	19	2	✘	N/A	2014

Cervical Cancer (Risk Factors) Data Set [25] (CCRFDS) comprises demographic information, habits, and historic medical records of 858 patients, with some missing attribute values due to the consents of the participating patients. The data set is categorised by the Boolean value of the biopsy. CCRFDS is also highly class-imbalanced, *i.e.* only 18 out of 858 participants have cancer, demonstrated in Fig. 4(a). Concretely, without including the class label, the original CCRFDS consists of 858 instances, each of which containing 35 attributes. However, there are 26 attributes containing missing values '?' due to the purpose of privacy protection. Thus, without the inclusion of the class label, we use 858 instances, each of which containing 9 attributes.

Breast Cancer Coimbra Data Set [50] (BCCDS) is composed of 116 data instances (each of which containing 9 attributes) that can be grouped into two categories, *i.e.* healthy controls, and patients.

Breast Tissue Data Set [51] (BTDS) contains 106 data instances, each of which with 9 feature dimensions that can be classified into six categories, including carcinoma, fibro-adenoma, mastopathy, glandular, connective, and adipose.

Diabetic Retinopathy Debrecen Data Set [52] (DRDDS) includes 1,151 data instances that are categorised into two classes which respectively indicate having and not having Diabetic Retinopathy (DR).
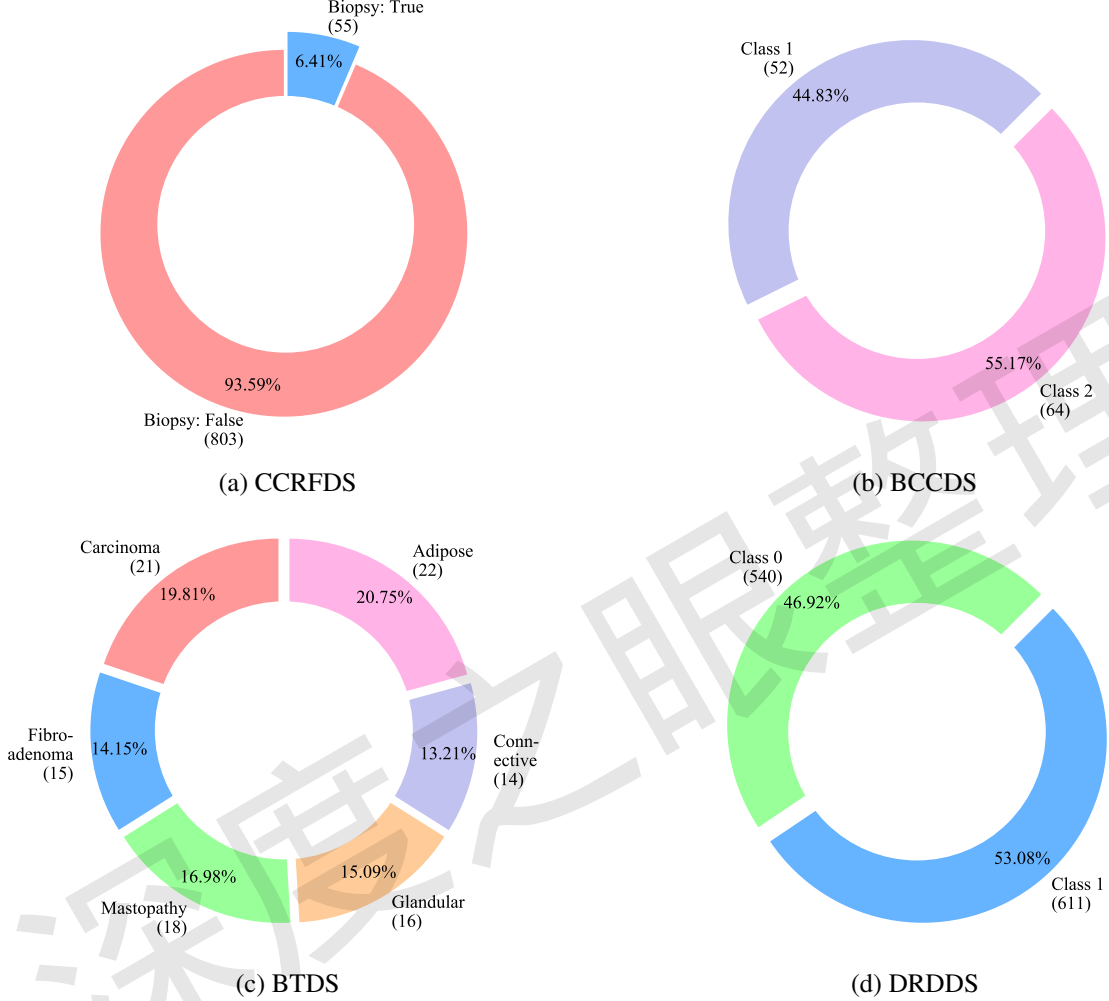

Figure 4: Data instances percentage for each data set. Best viewed in colour.

4.2 Dealing with Missing Data

Missing data, *a.k.a.* missing values, is a common issue in the digital healthcare domain. As introduced above, missing data could reduce the statistical power of the predictive model, as well as lead to incorrect or invalid results. Therefore, an extra stage may be required for handling the issue of missing data. There are two major methods that could be adopted to cope with this issue, *i.e.* data imputation and case deletion [53]. Concretely, imputation is the process of inserting the missed values with substituting components. Several approaches have been well discussed in the literature, such as mean, Multiple Imputation with Chained Equations-Full (MICE-full), and missForset [54]. Among those methods, the mean imputation approach imputes missing values as the mean of the available values for each variable, MICE-full and missForset then use Machine Learning algorithms, *e.g.* Random Forest, to predict the missing values based on the observed values of a data matrix. For the latter, all data instances with missing data are simply omitted/removed, and then only the remaining data is used for the analysis. Alternatively, the remaining data could be retained by removing the attributes in which the missing values are situated. In this work, we merely apply the attribute deletion method on CCRFDS.

4.3 Experiment Setup

Among all the four selected data sets, we perform data cleaning on CCRFDS, *i.e.* the attributes, which contain the missing value, are eliminated. Therefore, the CCRFDS used in this work contains 858 data instances, each of which having 9 attributes (not including the class label). For all the other three data sets, we use the originally published data.

For selecting features, we compare the proposed CFS using the Top$K$ method defined in Eq. (5) with the well-known PCA, IG, MI, and CST by varying the numbers of the selected features for classification. That is, we select 7 out of 9 attributes (in CCRFDS, BCCDS, and BTDS) and 15 out of 19 attributes (in DRDDS). To normalise the selected features, 8 normalisation methods introduced in Section 3.3 are employed, in which the power coefficient in PN and its variants are set to 0.1. For classification, 9 classifiers (introduced in Section 3.4) are employed with the configuration information: the maximum number of estimators is 100 at which the boosting is terminated in AB; the $\ell 1$ regularisation is adopted as the penalty function in LR; the Gini index is employed as the criterion in DT and the maximum depth of the tree is valued as 5; the number of neurons in a single hidden layer of BPNN is set to 20; the k is valued as 3 in kNN. The mean accuracy is reported for the nine employed classifiers for performance comparisons via the 10-Fold cross-validation.

4.4 The Efficacy of CFS

The detailed comparisons among the CFS and four counterparts are shown in Fig. 5 with Top Mean Accuracy (TMA) highlighted using different colour codes. In addition, those accuracies are further summarised in the first column of Fig. 6. In general, CFS outperforms PCA, IG, MI, and CST on the four data sets. Concretely, CFS yields averaged mean accuracies of 95.27%, 63.02%, 68.15%, and 65.61%, versus 94.87%, 60.65%, 52.27%, and 65.37% resulting from PCA; 94.96%, 61.20%, 65.06%, 64.59% generated by IG; 95.22%, 62.40%, 61.90%, and 64.44% produced by MI; and 95.14%, 62.82%, 62.62%, and 64.50% yielded by CST. This observation indicates that the CFS is generally more competitive than PCA, IG, MI, and CST.

For the CCRFDS, as depicted in Fig. 5(a), the TMA yielded by the CFS-based, PCA-based, IG-based, MI-based, and CST-based are 97.09%, 96.27%, 97.67%, 97.67%, and 96.39%, respectively. Among those best results, the MM normalisation approach is more suitable for PCA, MI, and CST, whereas the peak performance of CFS and IG are yielded by utilising PN$\ell 1$ and $\ell 1$PN normalisation approaches, respectively.

For the BCCDS, as depicted in Fig. 5(b), the TMA obtained by the CFS-based, PCA-based, IG-based, MI-based, and CST-based are 85.00%, 76.67%, 77.50%, 73.94%, and 75.83%, respectively. The conventional data normalisation approach tends to be an appropriate choice for CFS (MM), PCA (MM), and CST ($\ell 2$), whereas the variants of PN are more preferred by IG ($\ell 2$PN) and MI ($\ell 1$PN).

For the BTDS, as depicted in Fig. 5(c), we show that all feature selection methods achieve TMA using the MM normalisation approach. Furthermore, all the experimented filter feature selection methods (*i.e.* CFS, IG, MI, and CST) significantly outperform the feature extraction one (*i.e.* PCA). Among the feature selection approaches, our CFS is confirmed as having the highest degree of stability in terms of 68.15% averaged mean accuracy. On this basis, CFS is capable of helping to differentiate between certain categories of healthy or pathological breast tissue.

For the DRDDS, as depicted in Fig. 5(d), TMA of all the experimented methods are all yielded by using conventional feature normalisation approaches, *i.e.* MM (IG, MI, and CST), $\ell 1$ (PCA), and $\ell 2$ (CFS). Specifically, in contrast to BTDS, PCA (76.98%) achieves better TMA than all the filter feature selection methods, though on average CFS (65.61%) is the best.

To better explain the reason that CFS is a more competitive candidate of feature selection in comparison with PCA and the other three filter feature selection methods, we visualised the ranking of all the attributes (*i.e.* feature importance ranking) generated by CFS in the right column of Fig. 6.

4.5 The Efficiency of CFS-TSK+

Having the representative feature extraction (PCA) and feature selection (CFS) method, we detail the peak performances (TMA) of PCA-TSK+ and CFS-TSK+ in Table 2. Notably, in conjunction with Fig. 5, CFS-TSK+ achieved the best performance in the data sets of CCRFDS, BCCDS, and DRDDS. This observation confirmed the practicability and efficiency of combining the CFS with TSK+.

However, the best performances of CFS-TSK+ and PCA-TSK+ on BTDS help us to identify the possible drawback of the TSK+ in coping with classification tasks. That is, TSK+ is not sensitive to formulate a class boundary when the given data samples are sparsely distributed in the feature space. Alternatively, the rule base is not generalised well in the step of clustering where each cluster is corresponding to a fuzzy rule. Based on the time consumption, we did
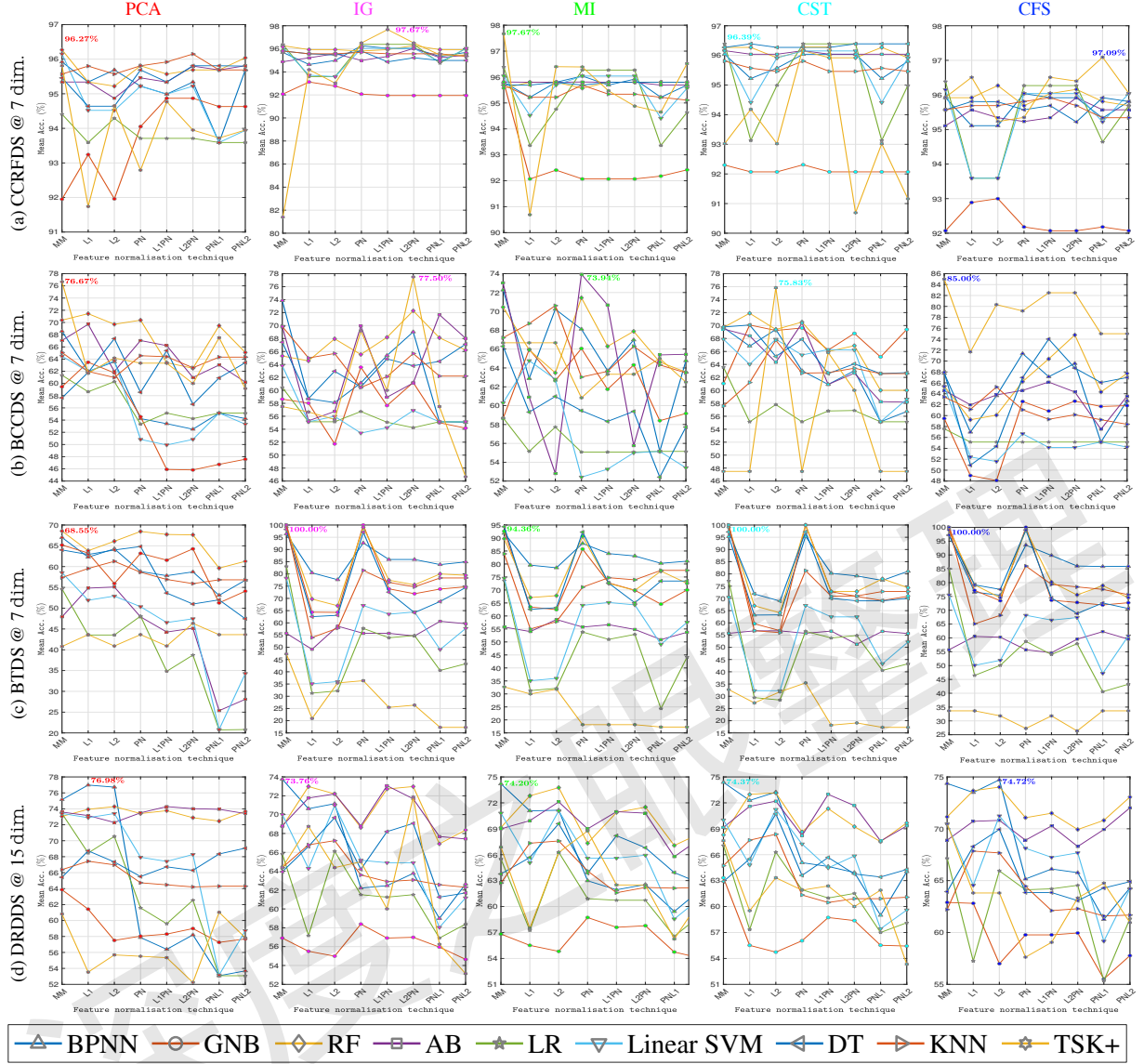
Figure 5: Detailed classification accuracies of PCA, IG, MI, CST, and CFS feature selection methods on four EHR data sets by varying both feature normalisation methods and classifiers. Best viewed in colour.

Table 2: Optimal performance comparisons of PCA-TSK+ and CFS-TSK+. Rule base employed in PCA-TSK+ and CFS-TSK+ were not optimised. ($\cdots$) denotes multiple combinations achieving the same performance. Best performance is marked in bold.

Data Set	PCA-TSK+ (%)↑	CFS-TSK+ (%)↑
CCRFDS	95.81 (MM)	**97.09** (PNℓ1)
BCCDS	76.67 (MM)	**85.00** (MM)
BTDS	**46.36** (ℓ2PN)	33.65 ($\cdots$)
DRDDS	61.03 (PNℓ1)	**70.48** (MM)%

not perform rule base optimisation in this work as this was slightly beyond our scope. For the last data set, DRDDS, owing greatly to the lack of expert knowledge, it is not possible to explain how reasonable the ranked results of CFS in comparison with the rest of the data sets are, which are more common sensible. The task of designing a self-explainable component could be treated as active future work.
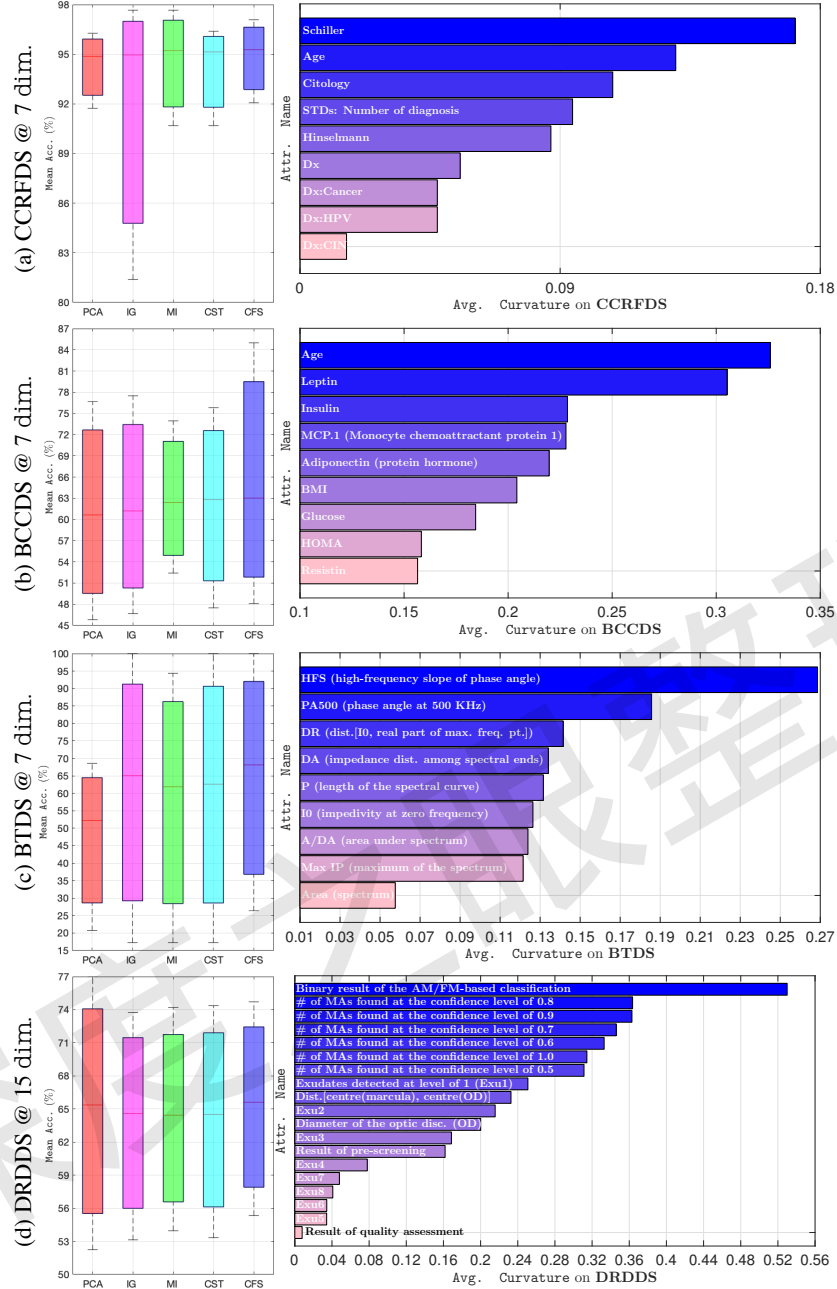
Figure 6: Classification summaries of PCA, IG, MI, CST, and CFS on four EHR data sets by varying both feature normalisation methods and classifiers. The left column summarises the classification statistics presented in Fig. 5, and the right columns lists the feature rankings provided by CFS in a descending order over each data set for intuitive explainability. Best viewed in colour.

4.6 Discussions

To summarise the proposed approach, we compare our CFS with PCA, IG, MI, CST and other recent competitive works in Table 3.

Though CFS achieved the best performance among all the four medical data sets, and CFS-TSK+ yielded the two highest mean accuracies on two data sets, we identified that a possible drawback of the proposed CFS is the lack of better explainability when the domain (*e.g.* clinical science) knowledge is not available. This might be mitigated by predicting the missing values on the anonymised data set and training a self-explainable component. In addition, a

Table 3: Performance summary and comparisons. DS denotes data set and TMA represents Top `Mean Acc.`. Abbreviations of all the experimented data sets are summarised in Table 1. $\cdots$ represents multiple combinations achieved the same performance (see Fig. 5(c). Best performance is marked in bold.

DS	Method	Dim.↓	TMA (%)↑
CCRFDS	BPCM+5NN [17]	9	80.00
	GSAM [18]	9	83.16
	PCA-RF	7	96.27
	CST-DT	7	96.39
	CFS-TSK+	7	97.09
	IG/MI-TSK+	7	**97.67**
BCCDS	WCNN [19]	9	62.00
	BLS II-SVM [20]	9	71.20
	MI-AB	7	73.94
	CST-TSK+	7	75.83
	PCA-TSK+	7	76.67
	IG-TSK+	7	77.50
	CFS-TSK+	7	**85.00**
BTDS	VIBES [21]	9	65.09
	PCA-RF	7	68.55
	2L-MJFA [22]	9	75.27
	MI-DT	7	94.36
	CFS/IG/CST-$\cdots$	7	**100.00**
DRDDS	RFGA [23]	19	68.26
	VAF [24]	19	73.35
	IG-BPNN	15	73.76
	MI-BPNN	15	74.20
	CST-BPNN	15	74.37
	CFS-BPNN	15	74.72
	PCA-BPNN	15	**76.98**

number of curvature-based methods, *e.g.* Integral Menger Curvature (IMC) and Gaussian Curvature (GC), can also be adopted to determine the curvature values of given data sets. Among such methods, Gaussian Curvature is usually applied to analyse the curvature of folded surfaces [55]. And the Integral Menger Curvature is the integral of the Menger Curvature over the whole given surface, which is related to the surface's Euler characteristic. The proposed method considers the curvature value of three discrete points. For this reason, GC and IMC is not implemented; however, it is still worthwhile to investigate how GC and IMC can support this approach in the future. Finally, four EHR data sets have been employed for the system evaluations; more complex data sets, such as those including labelled noise, would be applied to investigate the performance of the proposed technique when dealing with more complex situations.

5   Conclusion

A novel filter-based feature selection method, CFS, has been presented as contributing to the classification performance of clinical (EHR) data sets. State-of-the-art performance on all four benchmark clinical data sets has been achieved by the proposed CFS approach. Though lacking expert knowledge of clinical science, we visualise the results of feature ranking given by CFS to support better explainability. It is noteworthy that the self-explainability of CFS and sparsity awareness of class boundaries of CFS-TSK+ are observed as possible future directions. In addition, a given threshold is required by the proposed CFS approach for selecting the ranked features. It would be worthwhile to investigate how this threshold can be intelligently determined in the future. Also, due to the limitation of the filter method, the selected features may not be relevant to the class labels. It is interesting to introduce some learning approaches, such as [56], to consider the relationship between features and class labels while ranking the features for the MLaaS [57].

Acknowledgements

References

[1] M. Abdel-Basset, V. Chang, and N. A. Nabeeh. An intelligent framework using disruptive technologies for covid-19 analysis. *Technol. Forecasting Social Change*, page 120431, 2020.

[2] V. Chang. Computational intelligence for medical imaging simulations. *J. Med. Syst.*, 42(1):1–12, 2018.

[3] V. Chang. Data analytics and visualization for inspecting cancers and genes. *Multimedia Tools Appl.*, 77(14):17693–17707, 2018.

[4] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren. Deepigeos: A deep interactive geodesic framework for medical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(7):1559–1572, 2019.

[5] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu. Feature selection: A data perspective. *ACM Comput. Surv.*, 50(6):1–45, 2017.

[6] H. Duanmu, P. B. Huang, S. Brahmavar, S. Lin, T. Ren, J. Kong, F. Wang, and T. Q. Duong. Prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer using deep learning with integrative imaging, molecular and demographic data. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pages 242–252. Springer, 2020.

[7] C. Rietschel, J. Yoon, and M. V. D. Schaar. Feature selection for survival analysis with competing risks using deep learning. In *Proc. NeurIPS Work. Mach. Learn. Health*, 2018.

[8] S. Denaxas, P. Stenetorp, S. Riedel, M. Pikoula, R. Dobson, and H. Hemingway. Application of clinical concept embeddings for heart failure prediction in uk ehr data. In *Proc. NeurIPS Work. Mach. Learn. Health*, 2018.

[9] X. Chen, G. Yuan, W. Wang, F. Nie, X. Chang, and J. Z. Huang. Local adaptive projection framework for feature selection of labeled and unlabeled data. *IEEE Trans. Neural Netw. Learn. Syst.*, 29(12):6362–6373, 2018.

[10] D. Han and J. Kim. Unified simultaneous clustering and feature selection for unlabeled and labeled data. *IEEE Trans. Neural Netw. Learn. Syst.*, 29(12):6083–6098, 2018.

[11] Y. Liu, F. Tang, and Z. Zeng. Feature selection based on dependency margin. *IEEE Trans. Cybern*, 45(6):1209–1221, 2014.

[12] E. B. Mazomenos, D. Biswas, A. Acharyya, T. Chen, K. Maharatna, J. Rosengarten, J. Morgan, and N. Curzen. A low-complexity ecg feature extraction algorithm for mobile healthcare applications. *IEEE J. Biomed. Health. Inf.*, 17(2):459–469, 2013.

[13] S. T. Dumais. Latent semantic analysis. *Annu. Rev. Inf. Sci. Technol.*, 38(1):188–230, 2004.

[14] B. Lévy, S. Petitjean, N. Ray, and J. Maillot. Least squares conformal maps for automatic texture atlas generation. *ACM Trans. Graphics*, 21(3):362–371, 2002.

[15] R. Jensen and N. M. Parthaláin. Towards scalable fuzzy–rough feature selection. *Inf. Sci.*, 323:1 – 15, 2015.

[16] M. Christ, A. W. Kempa-Liehr, and M. Feindt. Distributed and parallel time series feature extraction for industrial big data applications. *arXiv preprint arXiv:1610.07717*, 2016.

[17] F.-Q. Li, S.-L. Wang, and G.-S. Liu. A bayesian possibilistic c-means clustering approach for cervical cancer screening. *Inf. Sci.*, 501:495–510, 2019.

[18] J. Lu, E. Song, A. Ghoneim, and M. Alrashoud. Machine learning for assisting cervical cancer diagnosis: An ensemble approach. *Future Gener. Comput. Syst.*, 2020.

[19] I. E. Livieris. Improving the classification efficiency of an ann utilizing a new training methodology. *Inf.*, 6(1):1, 2019.

[20] H. Tang, P. Dong, and Y. Shi. A construction of robust representations for small data sets using broad learning system. *IEEE Trans. Syst., Man, Cybern. Syst.*, 2019.

[21] F. Aydın and Z. Aslan. The construction of a majority-voting ensemble based on the interrelation and amount of information of features. *The Comput. J.*, 2019.

[22] X. Yang, K. Huang, R. Zhang, J. Y. Goulermas, and A. Hussain. A new two-layer mixture of factor analyzers with joint factor loading model for the classification of small dataset problems. *Neurocomputing*, 312:352–363, 2018.

[23] E. Elyan and M. M. Gaber. A genetic algorithm approach to optimising random forests applied to class engineered data. *Inf. Sci.*, 384:220–234, 2017.

[24] A. Apicella, F. Isgrò, and R. Prevete. A simple and efficient architecture for trainable activation functions. *Neurocomputing*, 370:1–15, 2019.

[25] K. Fernandes, J. S. Cardoso, and J. Fernandes. Transfer learning with partial observability applied to cervical cancer screening. In *Proc. Iberian Conf. Pattern Recognit Image Anal.*, pages 243–250, 2017.

[26] J. Li, Y. Qu, H. P. H. Shum, and L. Yang. Tsk inference with sparse rule bases. In *Adv. Comput. Intell. Syst.*, pages 107–123. Springer, 2017.

[27] J Li, L Yang, Y Qu, and G Sexton. An extended takagi–sugeno–kang inference system (tsk+) with fuzzy interpolation and its rule base generation. *Soft Computing*, 22(10):3155–3170, 2018.

[28] A. Ghoneim, G. Muhammad, and M. S. Hossain. Cervical cancer classification using convolutional neural networks and extreme learning machines. *Future Gener. Comput. Syst.*, 102:643–649, 2020.

[29] D. Devarriya, C. Gulati, V. Mansharamani, A. Sakalle, and A. Bhardwaj. Unbalanced breast cancer data classification using novel fitness functions in genetic programming. *Expert Syst. Appl.*, 140:112866, 2020.

[30] Z. Li, C. Wang, M. Han, Y. Xue, W. Wei, L.-J. Li, and L. Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 8290–8299, 2018.

[31] K. El Emam, S. Rodgers, and B. Malin. Anonymising and sharing individual patient data. *BMJ*, 350:h1139, 2015.

[32] C. L. Thompson and H. M. Morgan. Ethical barriers to artificial intelligence in the national health service, united kingdom of great britain and northern ireland. *Bull. World Health Organ.*, 98(4):293, 2020.

[33] V. Chang. An ethical framework for big data and smart cities. *Technol. Forecasting Social Change*, 165:120559, 2021.

[34] Z. Zuo, J. Li, B. Wei, L. Yang, F. Chao, and N. Naik. Adaptive activation function generation for artificial neural networks through fuzzy inference with application in grooming text categorisation. In *Proc. IEEE Int. Conf. Fuzzy Syst.*, pages 1–6, 2019.

[35] B. Remeseiro and V. Bolon-Canedo. A review of feature selection methods in medical applications. *Computers in biology and medicine*, 112:103375, 2019.

[36] Z. Zuo, L. Yang, Y. Peng, F. Chao, and Y. Qu. Gaze-informed egocentric action recognition for memory aid systems. *IEEE Access*, 6:12894–12904, 2018.

[37] N. A. Moubayed, B. A. S. Hasan, and A. S. McGough. Enhanced detection of movement onset in eeg through deep oversampling. In *IEEE Int. Jt. Conf. Neural Networks*, pages 71–78. IEEE, 2017.

[38] R. Aziz, C. K. Verma, and N. Srivastava. Dimension reduction methods for microarray data: a review. *AIMS Bioengineering*, 4(2):179–197, 2017.

[39] J. Li, Y. Qu, F. Chao, H. P. H. Shum, E. S. L. Ho, and L. Yang. Machine learning algorithms for network intrusion detection. In *AI in Cybersecurity*, pages 151–179. Springer, 2019.

[40] Z. Zuo, J. Li, P. Anderson, L. Yang, and N. Naik. Grooming detection using fuzzy-rough feature selection and text classification. In *Proc. IEEE Int. Conf. Fuzzy Syst.*, pages 1–8, 2018.

[41] H. Liang, X. Sun, Y. Sun, and Y. Gao. Text feature extraction based on deep learning: a review. *EURASIP J. Wireless Commun. Networking*, 2017(1):1–12, 2017.

[42] B. Azhagusundari, A. S. Thanamani, et al. Feature selection based on information gain. *Int. J. Innovative Technol. Exploring Eng.*, 2(2):18–21, 2013.

[43] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery, and N. Yazdani. Mutual information-based feature selection for intrusion detection systems. *J Network Comput. Appl.*, 34(4):1184–1199, 2011.

[44] J. Pohjalainen, O. Räsänen, and S. Kadioglu. Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. *Comput. Speech Lang.*, 29(1):145–171, 2015.

[45] P. Saengsiri, P. Meesad, S. N. Wichian, and U. Herwig. Comparison of hybrid feature selection models on gene expression data. In *Proc. IEEE Int. Conf. ICT Knowl. Eng.*, pages 13–18, 2010.

[46] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 13(5):971–989, 2015.

[47] D. Jain and V. Singh. Feature selection and classification systems for chronic disease prediction: A review. *Egyptian Info. J.*, 19(3):179–189, 2018.

[48] J.-C. Léger. Menger curvature and rectifiability. *Ann. of Math.*, 149:831–869, 1999.

[49] Z. Zuo, J. Li, and L. Yang. Curvature-based sparse rule base generation for fuzzy interpolation using menger curvature. In *Adv. Comput. Intell. Syst.*, pages 53–65, Cham, 2020. Springer.

[50] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça, and F. Caramelo. Using resistin, glucose, age and bmi to predict the presence of breast cancer. *BMC cancer*, 18(1):29, 2018.

[51] J. E. Da Silva, J. P. M. De Sá, and J. Jossinet. Classification of breast tissue by electrical impedance spectroscopy. *Med. Biol. Eng. Comput.*, 38(1):26–30, 2000.

[52] B. Antal and A. Hajdu. An ensemble-based system for automatic screening of diabetic retinopathy. *Knowledge-based Syst.*, 60:20–27, 2014.

[53] H. Kang. The prevention and handling of the missing data. *Korean J. Anesthesiology*, 64(5):402, 2013.

[54] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron. Using machine learning to predict laboratory test results. *Am. J. Clin. Pathol.*, 145(6):778–788, 2016.

[55] J. L. Richard. Detection of zones of abnormal strains in structures using gaussian curvature analysis. *AAPG Bull.*, 78(12):1811–1819, 1994.

[56] J. Li, H. P. H. Shum, X. Fu, G. Sexton, and L. Yang. Experience-based rule base generation and adaptation for fuzzy interpolation. In *Proc. IEEE Int. Conf. Fuzzy Syst.*, pages 102–109, 2016.

[57] V. Chang. An overview, examples, and impacts offered by emerging services and analytics in cloud computing virtual reality. *Neural Comput. Appl.*, 29(5):1243–1256, 2018.