

Research paper

Evolutionary deep feature selection for compact representation of gigapixel images in digital pathology

Azam Asilian Bidgoli^a, Shahryar Rahnamayan^{b,c,*}, Taher Dehkharghanian^d, Abtin Riasatian^c, Shivam Kalra^c, Manit Zaveri^c, Clinton J.V. Campbell^d, Anil Parwani^e, Liron Pantanowitz^f, H.R. Tizhoosh^{c,g}

^a NICI Lab¹, Ontario Tech University, Oshawa, Canada

^b NICI Lab², Brock University, St. Catharines, Canada

^c Kimia Lab³, University of Waterloo, Waterloo, Canada

^d Department of Pathology and Molecular Medicine, McMaster University, Hamilton, Canada

^e Department of Pathology, Wexner Medical Center, Ohio State University, Columbus, USA

^f Department of Pathology, University of Michigan, Ann Arbor, USA

^g Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, USA

ARTICLE INFO

Keywords:

Digital pathology

Whole slide images

Image representation

Evolutionary computation

ABSTRACT

Despite the recent progress in Deep Neural Networks (DNNs) to characterize histopathology images, compactly representing a gigapixel whole-slide image (WSI) via salient features to enable computational pathology is still an urgent need and a significant challenge. In this paper, we propose a novel WSI characterization approach to represent, search and classify biopsy specimens using a compact feature vector (CFV) extracted from a multitude of deep feature vectors. Since the non-optimal design and training of deep networks may result in many irrelevant and redundant features and also cause computational bottlenecks, we proposed a low-cost stochastic method to optimize the output of pre-trained deep networks using evolutionary algorithms to generate a very small set of features to accurately represent each tissue/biopsy. The performance of the proposed method has been assessed using WSIs from the publicly available TCGA image data. In addition to acquiring a very compact representation (i.e., 11,000 times smaller than the initial set of features), the optimized features achieved 93% classification accuracy resulting in 11% improvement compared to the published benchmarks. The experimental results reveal that the proposed method can reliably select salient features of the biopsy sample. Furthermore, the proposed approach holds the potential to immensely facilitate the adoption of digital pathology by enabling a new generation of WSI representation for efficient storage and more user-friendly visualization.

1. Introduction

The availability of whole slide images (WSIs) in digital pathology has provided researchers with new opportunities to investigate tissue samples using computational algorithms. Machine learning in general and deep learning in particular are increasingly influencing digital pathology workflow [1]. Machine learning models are required to be trained by sufficiently large and diverse annotated (labeled) datasets. However, publicly available medical image datasets usually lack these two characteristics [2]. For instance, The Cancer Genome Atlas (TCGA) repository [3] only contains around 33,000 unlabeled WSIs of 32

cancer types while *ImageNet*, an image database of common objects, is composed of more than 14 million images [4]. Furthermore, medical image datasets only include a handful of disease types and are often not annotated at the pixel level, which is another obstacle that hinders the implementation of classification models for clinical utility [2].

Image search is one of the promising applications of computerized solutions in digital pathology [5–8]. It enables pathologists to query patient images or a particular region of interest to find images with similar histomorphological features within a large WSI archive. Through the information obtained from retrieved similar cases, end-users can hone their initial diagnosis, get a better understanding of the

* Corresponding author at: NICI Lab, Ontario Tech University, Oshawa, Canada.

E-mail address: shahryar.rahnamayan@uoit.ca (S. Rahnamayan).

¹ Nature Inspired Computational Intelligence Lab.

² Nature Inspired Computational Intelligence Lab.

³ Laboratory for Knowledge Inference in Medical Image Analysis.

patient's prognosis, and eventually devise an accurate final diagnostic interpretation when integrated with ancillary data. Therefore, image search can contribute towards precision medicine [9].

The first step of any image search process is extracting features to represent an image. Ideally, these extracted features should semantically reflect the image content rather than just raw inexpressive pixel information. Analogous to other computer vision fields, image retrieval has also been heavily influenced by the emergence of Convolutional Neural Networks (CNNs). Today, many researchers prefer to use pre-trained CNNs for feature extraction [6,7,10]. Generally speaking, WSIs have extremely large dimensions; therefore, it is very challenging to feed them directly into CNNs at high magnification due to computational and memory constraints. Hence, smaller sized image tiles or patches should be extracted from different WSI regions to represent its content. However, a large number of tissue patches, the length of deep feature vectors, and sophisticated methods to represent WSIs make it difficult to efficiently perform and interpret any downstream task including image search.

Due to the large number of feature vectors representing a typical histopathology image, expensive computations for image search are required. In order to find the most similar images in a dataset with millions of images, the distance between every feature vector of the query image and all feature vectors of all archived images should be calculated. Using hashing and converting the real-valued features to binary ones could alleviate the heavy computation of the image retrieval [10–13]. However, binarization inevitably causes the loss of information. In contrast, eliminating redundant and irrelevant features can generate a compact (short) code which facilitates the digital pathology processes such as image retrieval. Furthermore, a smaller sized deep feature vector requires fewer memory bytes for storage. Hence, by reducing the length of DNN feature vectors, it would be possible to sample a larger number of tissue patches per WSI.

To facilitate the classification of gigapixel WSIs, some efforts are done in pixel-level compression. Tellez et al. [14] designed a framework to compress the images using a supervised approach before classification. In addition, the image can be compressed before processing. Helin et al. [15] defined a parameterization for WSI compression with different degree on background and tissue-containing parts. A Neural Image Compression is presented in [16] which compresses the WSIs using DNNs such as autoencoder to be processed using CNNs for classification. Despite the benefits of these methods, they just try to tackle the processing of histopathology images while the representative code of the image is still large.

A reduction in the length of deep feature vectors could improve visualization capability of image search engine. It is possible to attribute histomorphologic patterns to deep features of a DNN trained on brain cancer images [17]. Hence, by storing the compact feature vectors, we can not only search at lower cost, but we could succinctly explain why two WSIs are considered similar by comparing tissue patches of query and retrieved WSIs. This would allow pathologists to quickly evaluate retrieved images without the necessity of an exhaustive investigation of WSIs. Such visualization could potentially improve user acceptance of image search as it is quite exhaustive to investigate a WSI, considering the large number of slides produced every day [18].

Generally speaking, DNN-derived feature vectors used for image representation are usually of length 512 to 2048. On the other hand, due to the un-optimality of DNNs, there are many redundant and irrelevant features among extracted features. Therefore, feature selection not only decreases the size of representative code, but also increase the efficiency of features by removing the redundant and irrelevant features. Consequently, storing a large number of DNN-derived feature vectors is a substantial memory burden plus slow processing algorithms, which often make them impractical in real-world applications. Hence, by decreasing the length of a DNN-derived feature vector, it would be plausible to increase the number of DNN-derived feature vectors per WSI, which would eventually improve WSI representation accordingly.

In the light of these motivations, feature selection is one of the well-known techniques that can be employed to reduce the size of the DNN-derived feature vector accordingly. To the best of our knowledge, this is the first time that WSIs are presented by an extremely compact code, i.e., 13,800 times shorter than initial length compared to the state-of-the-art algorithms. In addition to compactness, the method finds an optimal feature vector to represent a WSI for accurate digital pathology applications such as image retrieval.

In this paper, an evolutionary method is presented to compress the feature vectors of histopathology images. The optimal features are selected among the deep features extracted from the images using a pre-trained DNN. The main contribution of this work is to generate a very compact code to represent a gigapixel histopathology image which is commonly represented by many deep feature vectors. The compact codes are generated based on the selected features specialized for each tumor type. Optimizing pre-trained features is meaningful since due to the lack of sufficient data, training a tailored DNN for each disease type may not be feasible.

2. Related works

The gigapixel size of histopathology images motivates researchers to study on finding novel approaches to decrease its size of representative vectors; otherwise, processing of them would be practically very complicated or even impossible. This is extremely crucial because the common way of representing a pathology image is to extract a large number of patches from a WSI, each of which requires a feature vector as its representative vector. In recent years, pre-trained DNNs have become the method of choice for feature extraction of histopathology images to be utilized in digital pathology including Content-based Image Retrieval (CBIR) [2,8,19]. Luigi, Yottixel, and SMILY are three examples of CBIR systems that use pre-trained DNNs for deep features extraction. Luigi is an application which uses deep texture representations created by a pre-trained DNN, and exploits the nearest neighbor search in order to retrieve similar cancer histopathology images [20]. Yottixel represents each WSI with a mosaic set of tissue patches. Then, deep features extracted from each tissue patch are converted into binary barcodes, which require lower size storage space [21]. SMILY applies a pre-trained DNN to tissue images at 10x magnification to build a search dataset using an embedding computational module [6]. All of the above-mentioned CBIR systems have utilized WSIs taken from TCGA repository [22].

Despite the useful employment of DNNs, the deep features are required to be efficiently optimized in many applications. Feature selection as a process of removing redundant and irrelevant features is one of the leading steps in pattern recognition and machine learning. In other words, the purpose of feature selection is to find a set of proper features so that the efficiency of the learning algorithm is improved. In many studies, it has been showed that the eliminating the irrelevant features enhances the performance of the learning models [23–25]. Finding the appropriate subset among N features requires evaluating of 2^N subsets for a given problem, which makes utilization of brute-force search impossible for large-scale cases. Additionally, feature selection effectively reduces the computational complexity of learning process. Although, feature selection can enhance the accuracy of the classification task and decrease the computational complexity, an extreme reduction of relevant features will degrade the accuracy [26]. Thus, feature selection can be modeled as an optimization problem with two conflicting objectives, maximizing the accuracy of classification and minimizing the number of features simultaneously, which is the main goal of this study to compress the representative code. Evolutionary Algorithms (EAs) as a category of optimization algorithms have been widely employed as the feature selectors in many machine-learning or data mining related applications [27–29].

Deep feature selection has been conducted in some recent successful studies. In [30], authors applied Kruskal–Wallis feature selection on a

set of deep features along with a set of classical hand-crafted features to select the best combination. The features are extracted from chest pathology images. The reduction of features is incurred using the minimum redundancy maximum relevance algorithm on the output of three deep models to select the best features set extracted from X-ray images [31]. Accordingly, the authors combined the selected features from independent deep models to provide an efficient feature set. In a similar study, Ozyurt developed a feature selection framework on deep features generated by several well-known pre-trained networks [32]. The features are selected based on the Relief algorithm. In [33], the deep features extracted by a teacher autoencoder are fed into a shallow student network to rank the low-dimensional deep features. At the last step of the proposed framework, the top ranked features are selected by the weights inspired from the student network. [34] characterizes structure and classification decisions using dimensionality reduction. Although they show the tile samples in 2-D space, the main purpose of the method is for visualization and outlier detection. Consequently, still the required data to represent the histopathology images is extremely large. A comprehensive comparative study was conducted by evaluating eleven feature selection algorithms on three conventional DNN algorithms, i.e., convolution neural network (CNN), deep belief network (DBN), recurrent neural network (RNN), and three recent DNNs, i.e., MobilenetV2, ShufflenetV2, and Squeezenet [35]. The experimental data supported their hypothesis that feature selection algorithms may improve deep neural network performances. Despite all studies on feature selection for representation of histopathology images, the providing an accurate compact code to characterize the images is a requisite for efficient digital pathology.

3. Proposed feature selection

3.1. Data and deep feature extraction

We used TCGA repository which consists of 32,072 WSIs for 32 primary diagnoses. The labeling of these images is at the WSI level (i.e., no pixel-level delineations) and includes information such as 'morphology', 'primary diagnosis' and 'tissue or organ of origin'. The histopathology features are extracted from the images using a deep network proposed in [36]. The authors have extracted image patches of size 1000×1000 pixels at $20\times$ magnification with no overlap from each WSI so that the image is represented by 130 patches on average. The size of patches is established based on the input default size of KimiaNet which is served as extractor. The details of network design are according to [36] in which the authors believe that the size of 1000×1000 is the large enough size for patches to model and cover the workflow for most diagnostic cases from the images. Then, the proposed DNN is feed by patches to results in 1,024 features for each patch. Accordingly, a WSI requires $130 \times 1024 = 133,120$ values to be represented digitally. Similar to mentioned study, a tumor site categorization of 30 tumor diagnoses was established to 12 tumor sites according to [37] including brain, breast, endocrine, gastrointestinal tract, gynecological, hematopoietic, melanocytes, mesenchymal, liver/pancreaticobiliary, prostate/testis, pulmonary, and urinary tract. All tumor sites except breast, hematopoietic tissue, and mesenchymal consisted of more than one tumor diagnosis. Table 4 in the Appendix represents the defined identity (ID) for each primary diagnosis and the number of patients in the dataset.

In order to compare the efficiency of optimal features with the original features extracted by the DNNs, same data sampling is required. Accordingly, the validation and test datasets each comprised of 10% of the whole dataset and were chosen randomly from the cases with a single WSI. This resulted in 741 and 744 WSIs for the validation and test datasets, respectively. The remaining 7,126 slides formed the training dataset.

3.2. Cross-patch average feature pooling

Thus far we used a common method to represent a pathology image by extracting a small number of patches from the WSI, in our case 15%

of all tissue patches. Although this strategy resulted in a practical way to break a WSI down into a more manageable format, multiple feature vectors corresponding to extracted patches are a serious obstacle to design practical diagnostic workflows in digital pathology. We accordingly propose to *average* all feature vectors of all patches into *one single feature vector* for each WSI. Average pooling has been subject to investigations for a long time [38]. But pooling usually happens inside the network. We propose to apply average feature pooling after extraction. The importance of average, i.e., central tendency, is obviously stated in statistics. The average uses every value in the data and hence is a good representative of the data.

On the other hand, averaging as a collaboration strategy is widely used in different applications such as model averaging on ensemble machine learning [39] or federated learning [40]. Moreover, as the trained DNN are operating on hypercellular feature vectors, average pooling is expected to provide a reliable representation for the entire WSI. For this purpose, the mean feature vector (MFV) across all patches is calculated to be the representative of the corresponding extracted features. Fig. 1 illustrates the process used to generate the MFV. A major criterion for patch selection in DNN training was the high value of cellularity. Cellularity can be used as an initial filter to select patches with a higher probability of malignancy. In fact, cellularity of patches can be used to eliminate most benign/healthy patches because carcinomas are generally associated with uncontrolled cell growth. Therefore, averaging feature values of such patches with high similarity is quite meaningful to generate a representative for the corresponding WSI. Having MFV as one feature vectors of length 1,024 is a considerable reduction of WSI feature space. Assuming a WSI is composed of 130 high cellular patches (on average) so that $130 \times 1,024 = 133,120$ real-valued numbers are generated for one WSI; using MFV, this large number decreases to only 1024 features (i.e., 130 times smaller). But there is a possibility to remove the redundant and irrelevant features for each tumor type. We have to smartly select $N \ll 1,024$ features out of MFV such that we can conveniently visualize them.

3.3. Evolutionary feature selection

MFV may still contain many redundant and potentially irrelevant features due to imperfections of training. Feature selection is a crucial task in machine learning and data mining to choose a small set of salient features among a larger set of features. The aim of feature selection is not only to enhance the classification or search process but also to reduce the computational complexity [41]. There are a variety of techniques that can be applied and tailored for image analysis [30,42,43]. However, feature selection is mainly challenging due to the large search space, especially when the number of original features is high. With 1,024 features, the number of possibilities for selecting around 10 different features is $\approx 3.342 \times 10^{23}$. Consequently, the evaluation of all subset of features is practically impossible. Hence, an efficient optimization technique is required to find an optimal feature subset in such a large search space. Evolutionary computation (EC) [27] is a category of stochastic optimization algorithms to search for a subset of features using a suitable "fitness" function. The candidate feature subsets are evaluated during the optimization process – inspired by the natural evolution – driven by a desired measure (i.e., a fitness function) such as F1-score or accuracy with the aim of increasing the performance of an algorithm; after many generations of evolutionary operations, the fittest features survive that we call compact feature vector (CFV).

The main objective of image retrieval is to search for specific regions of a slide similar to the query case. Therefore, by transferring the knowledge from features extracted on images of all tumor subtypes to a compact subset of features, a more efficient image retrieval system will be proposed to discriminate between a number of tumor subtypes in a specific category.

In fact, the size of the extracted feature vector in a deep network is based on the designer's decision and may not be optimal for an entire

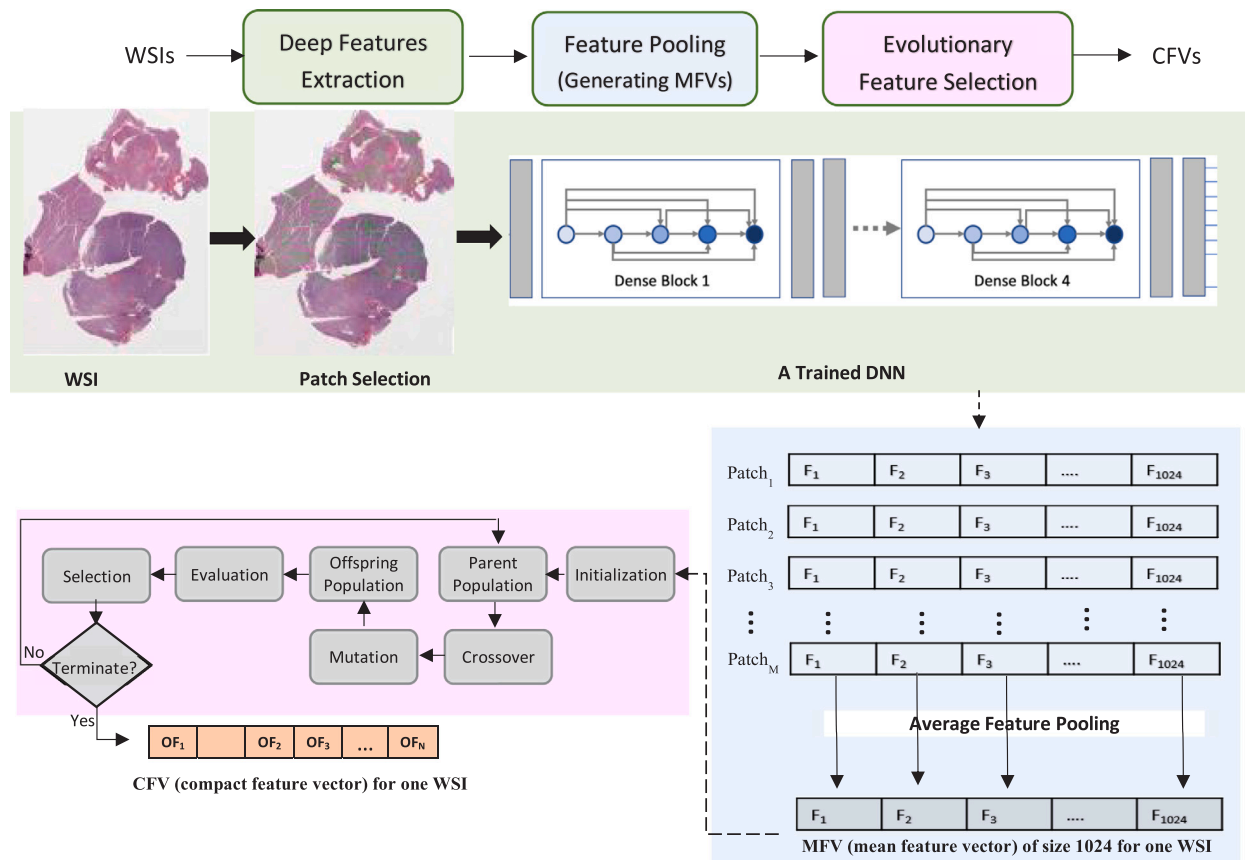


Fig. 1. Overview: The process of creating the mean feature vector (MFV) and compact feature vector (CFV). The deep feature extraction phase is conducted by a deep neural network trained on selected patches, which is proposed in [36]. A smart feature selection must then further shorten the MFV to a compact feature vector.

dataset or individual primary sites. For instance, trained DNN used a fix-length feature vector of 1,024 values for all tumor types whereas feature selection finds an optimal subset of features for each organ.

Evolutionary algorithms are a category of optimization methods that can search a huge space of possibilities to find the optimal or nearly optimal values for a set of variables according to minimizing or maximizing one (or multiple) objective functions. Accordingly, in a feature selection problem, the main goal is to select the best subset of features to boost an evaluation measure such as F1-score for a desired task (e.g., classification or search). In addition to increasing the accuracy, other objectives such as minimizing the number of selected features can be considered [44]. In the case of two objectives, a *multi-objective optimization* process is designed. Since in a multi-objective optimization problem different objectives may be in conflict, the definition of optimality is straightforward. Multi-objective approaches make a trade-off decision to select a set of solutions instead of offering only one solution. Therefore, new concepts are required to compare two candidate solutions and select the best one; the concept of “dominance” is one of the well-known relations for this purpose.

A population-based evolutionary algorithm works with a set of individuals as a population that are potentially candidate solutions for the optimization problem. Similar to evolution in nature, the population evolves to reach the best candidate solutions; accordingly, new individuals are created using generative operators such as crossover, mutation and parent selection, then some individuals survive and emerge as the new generation. For this purpose, a selection operator based on an evaluation measure decides which individuals should remain in the population. The process continues until a stopping criterion such as a predefined number of iterations, is met.

At the beginning, uniform random individuals are produced as an initial population. Feature selection is originally a high dimensional

binary optimization problem. Each individual in the population is a binary vector in the length of the number of features. Each variable of the vector represents the status of a feature so that presence or absence of a feature is indicated by 1 or 0, respectively. Each feature subset in the population is evaluated based on both objectives, number of features and the average of the F1-score [45,46] when tested on the primary diagnosis according to the ground truth labels provided in TCGA dataset. In order to optimize the space by selecting new feature subsets, evolutionary operators are applied on the population to generate new individuals. Crossover is an evolutionary operator to combine two candidate solutions (i.e., parents) to transfer a portion of information from each individual to offsprings. The employed combination method is one-point crossover [47] in which from a random point, the genes of both individuals are exchanged. By this way, the features of two selected subsets create new combinations of features. Additionally, the mutation operator as another evolutionary operator is applied to increase the diversity of the population. Bit-wise mutation [48] selects a portion of genes to invert their values so that some features are removed from the current set and others are added to the list. The next step of an evolutionary algorithm is the selection of the best candidate solutions from the combination of old population and offsprings to produce the next generation. In a multi-objective optimization problem, the selection method is applied based on the dominance concept. One of the well-known multi-objective algorithms is the third version of Non-dominated Sorting Genetic Algorithm (NSGA-III) [49] which ranks the individuals hierarchy based on the number of dominated points. The best subset of features are selected from the top ranked “Pareto fronts” to be included in the new generation.

In order to evaluate the quality of feature subsets to find CVF, we used the *k*-Nearest Neighbor (*k*-NN) approach [50] to find the *k* most similar images to a query WSI among the validation set and accordingly

calculate F1-score as the first objective. k -NN calculates the Euclidean distance between the query image and all images in the dataset. The query WSI is labeled by the primary diagnosis or tumor type which is most frequent among the k search results.

As mentioned previously, the second objective of the optimization problem to reduce the number of features. At the end of the algorithm, the objectives of the final solution subsets are reported on the test set. The overall structure of the evolutionary algorithm is illustrated in Fig. 8 in Appendix. In summary, the process of compactness of the image representation consists of two main steps: producing the mean feature vector from the original features extracted from the patches and then selection the optimal features among the mean features to represent the entire WSI using only a compact feature vector.

In the following, the effectiveness of both representation strategies are evaluated. Firstly, the mean features are assessed by employing them to accomplish a query WSI. Then, the search is conducted by compact feature vectors and correspondingly the confusion matrix and ROC plots of the obtained results are presented. Moreover, the proposed method is compared with the state-of-the-art dimension reduction methods to reveal the outperformance of the evolutionary-based features.

4. Experiments

In order to assess the compact feature vector (CFV) for each WSI, we conducted several experiments. The development and all experiments are conducted on publicly available TCGA image dataset. The histopathology images are divided to three subsets as training, validation and, test. Since we required to utilized the extracted features from [36] and consequently we should not have used their test dataset during the method development, the same subsets are used. All results are reported on the test dataset consisting of 744 WSIs comprised of 30 different primary tumor diagnoses.

After optimizing the feature vectors, the retrieval is accomplished among the test slides of a specific anatomic site (but containing different primary diagnoses). Each category consists of at most four primary diagnoses, and a true positive is considered when the correct primary diagnosis is predicted. The evolutionary algorithm selects the optimal subset of features to boost the image search results. Accordingly, for each tumor site category, an optimization problem is defined separately to select more relevant features to discriminate the tumor subtypes of the corresponding anatomic site. For our experiments, the entire dataset is divided to three subsets. The training and validation sets, which had been utilized for training and validation of DNN, are employed during the optimization process so that for each candidate feature subset, WSIs of the validation set are searched among those of the training set to find the most similar cases. Thus, the test set is remained for evaluation phase. The output of multi-objective optimization is a set of trade-off solutions (i.e., Pareto front). To assess the method on test set, the best feature subset with maximum average F1-score resulted from the optimization process (i.e., validation set) is selected and the evaluation measure is calculated on test set. Thus, the best subset is selected based on the training and validation, but the final result is reported on test set. k -nearest neighbor algorithm with $k = 3$ and Euclidean distance are utilized to find three of the most similar images, and consequently employed to calculate the average F1-score over all primary diagnosis as the first objective of the optimization problem. F1-score is one of the common measures to assess the search process which is a harmonic mean of precision and recall measures.

Due to the stochastic nature of the evolutionary algorithms, we run 31 independent runs of the algorithm to obtain results. Finally, we then evaluate the resulted features on “Pareto front” [49] by performing image search on the test dataset. Finally, the Wilcoxon statistical test [51] a 95% confidence level is conducted to investigate the significance of the acquired results.

4.1. Histopathology image retrieval using MFV

The search is conducted on 29 different primary diagnoses. A WSI is matched against the slides of the same tumor category to find three of the most similar images. The aim is to recognize the primary diagnosis of the query WSI. The results using the original features (i.e., on overage 130 feature vectors for each WSI) and MFV in terms of F1-scores are presented in Table 1. As the results indicate, despite compacting the features into one single MFV, in most categories the retrieval has become more precise. For instance, the discrimination between two tumor subtypes, LGG and GBM, is enhanced 4% compared to the result of the original features. The average improvement by MFVs of deep features is more than 2%.

Consequently, MFV performance surpasses the original features for image retrieval, providing a significantly compact WSI representation. The confusion matrix (Fig. 9) and t-SNE visualization (Fig. 10) of this experiment are provided in Appendix.

As mentioned previously, the main reason behind the efficiency of MFV is the utilization of high cellular regions of a WSI which leads to more reliable representative. As a result, the average of deep features extracted using a network which is trained on such selected patches can effectively replaces the entire feature vectors. Conversely, this benefit cannot be resulted in averaging of features extracted by a pre-trained network which is not trained by high cellular patches. To verify this point, Table 6 in appendix represents the result of averaging the deep features of the DenseNet as a pre-trained network for extracting the features from the histopathology images.

In addition to the presented search results (namely vertical search), an alternative search (i.e., horizontal search) in which a query WSI is searched among the entire test set to find the most similar images is conducted; a true positive match is incurred when the slides with the same tumor type are found. To evaluate the MFV on the horizontal search, the results of this experiment utilizing the features extracted by the mentioned DNN is provided in the Table 5 of Appendix.

4.2. Histopathology image retrieval using CFV

We previously showed that MFV as one feature vector can represent a WSI. In this next experiment, we intend to show that the compact feature vector (CFV), the evolutionary optimized MFV, can also represent a WSI for a correctly identified primary diagnoses. The results of the search by CFV are presented in Table 2. In addition to accuracy improvement, reduction in dimensionality of the code representing a WSI is a consequential benefit of feature selection. The average number of optimized features to represent a WSI is 12 feature values which is an astonishingly small number. The F1-score values on efficiency of two set of features including MFV and CFV are reported. For most of the groups, the selected subset of features in CFV surpasses the MFV in order to identify the primary diagnosis. For instance, two tumor types of the brain are separated by 97% of F1-score when using only 8 features. CFV is $1024/8 = 128$ times smaller than MFV. Compared to the original deep features with an average of 130 patches/WSI, CFV is $1024 * 130/8 = 16,640$ times smaller.

Subtype Visualization — In addition to numerical results, the t-SNE visualization resulted by the evolutionary features on each tumor site is illustrated in Fig. 2. The optimized feature vectors of each tumor type category are mapped into 2-dimensional space to represent the distinction between WSIs of each primary diagnosis. There is relevant concordance between the numerical results and visualization output with respect to the separation or overlap among different classes. For instance, according to Table 2 the resulted values of F1-score by CFV to discriminate the rectal adenocarcinoma (READ) from colon adenocarcinoma (COAD) are 76% and 47%, respectively. That is consistent with t-SNE visualization in which there is some overlapping between the samples of READ and COAD.

Table 1

The results of the search for all tumor subtypes. Each WSI of the test set is matched against the images of the same tumor type to identify the correct primary diagnosis. The F1-score is reported on three-nearest neighbor search using “median-of-min” approach on all features extracted from each image which contains on average 130 feature vectors of size 1024 and MFV which is the mean feature vector of size 1024.

Tumor type	Subtype	# WSI	All features (130 × 1024)	MFVs (1024)	Diff
Brain	LGG	35	81.08	84.51	+3.51
	GBM	39	81.08	85.71	+4.71
Endocrine	ACC	6	44.44	54.55	+10.55
	PCPG	15	84.85	83.87	-1.13
	THCA	51	100	100	0
Gastrointestinal	COAD	33	76.47	76.71	+0.71
	READ	11	30	42.11	+12.11
	ESCA	14	78.26	84.62	+6.62
	STAD	30	86.15	79.31	-6.69
Gynecologic	CESC	17	94.12	97.14	+3.14
	OV	10	94.74	94.74	-0.26
	UCS	3	85.71	100	+14
Liver	CHOL	4	40	50	+10
	LIHC	35	95.77	95.77	-0.23
	PAAD	12	76.19	73.68	-2.32
Melanocytic	UVM	4	66.67	66.67	-0.33
	SKCM	24	93.62	96	+2
Prostate	PRAD	40	100	100	0
	TGCT	13	100	100	0
Pulmonary	LUAD	43	78.33	81.58	+3.58
	LUSC	38	84.44	86.36	+2.36
	MESO	5	75	75	0
Urinary tract	BLCA	34	93.15	94.29	+1.29
	KICH	11	85.71	90	+4
	KIRC	50	96.97	95.05	-1.95
	KIRP	28	90.57	87.27	-3.73
Average			81.28	83.65	+2.38

Table 2

The results of the search by evolutionary features. The comparison between the efficiency of CFV and MFV is provided in terms of F1-scores by three-nearest neighbor approach. Whereas MFV has 1024 feature values, the optimized number of features, #F, for CFV, indicated for each site in the last column, is on average 12 features.

Tumor type	Subtype	# WSI	MFV (1024)	CFV (≈ 12)	Diff	#F
Brain	LGG	35	84.51	97.43	+12.92	8
	GBM	39	85.71	97.14	+11.43	
Endocrine	ACC	6	54.55	100	+45.45	2
	PCPG	15	83.87	100	+16.13	
	THCA	51	100	100	0	
Gastrointestinal	COAD	33	76.71	76.06	-0.66	28
	READ	11	42.11	47.62	+5.51	
	ESCA	14	84.62	84.62	0	
	STAD	30	79.31	89.66	+10.34	
Gynecologic	CESC	17	97.14	100	+2.86	14
	OV	10	94.74	100	+5.26	
	UCS	3	100	100	0	
Liver	CHOL	4	50	75	+25	11
	LIHC	35	95.77	97.14	+1.37	
	PAAD	12	73.68	100	+26.32	
Melanocytic	UVM	4	66.67	100	+33.33	6
	SKCM	24	96	100	+4	
Prostate	PRAD	40	100	100	0	1
	TGCT	13	100	100	0	
Pulmonary	LUAD	43	81.58	88.89	+7.31	14
	LUSC	38	86.36	91.11	+4.75	
	MESO	5	75	100	+25	
Urinary tract	BLCA	34	94.29	95.65	+1.37	21
	KICH	11	90	95.24	+5.24	
	KIRC	50	95.05	96.91	+1.86	
	KIRP	28	87.27	91.53	+4.25	
Average			83.65	93.23	+9.58	12

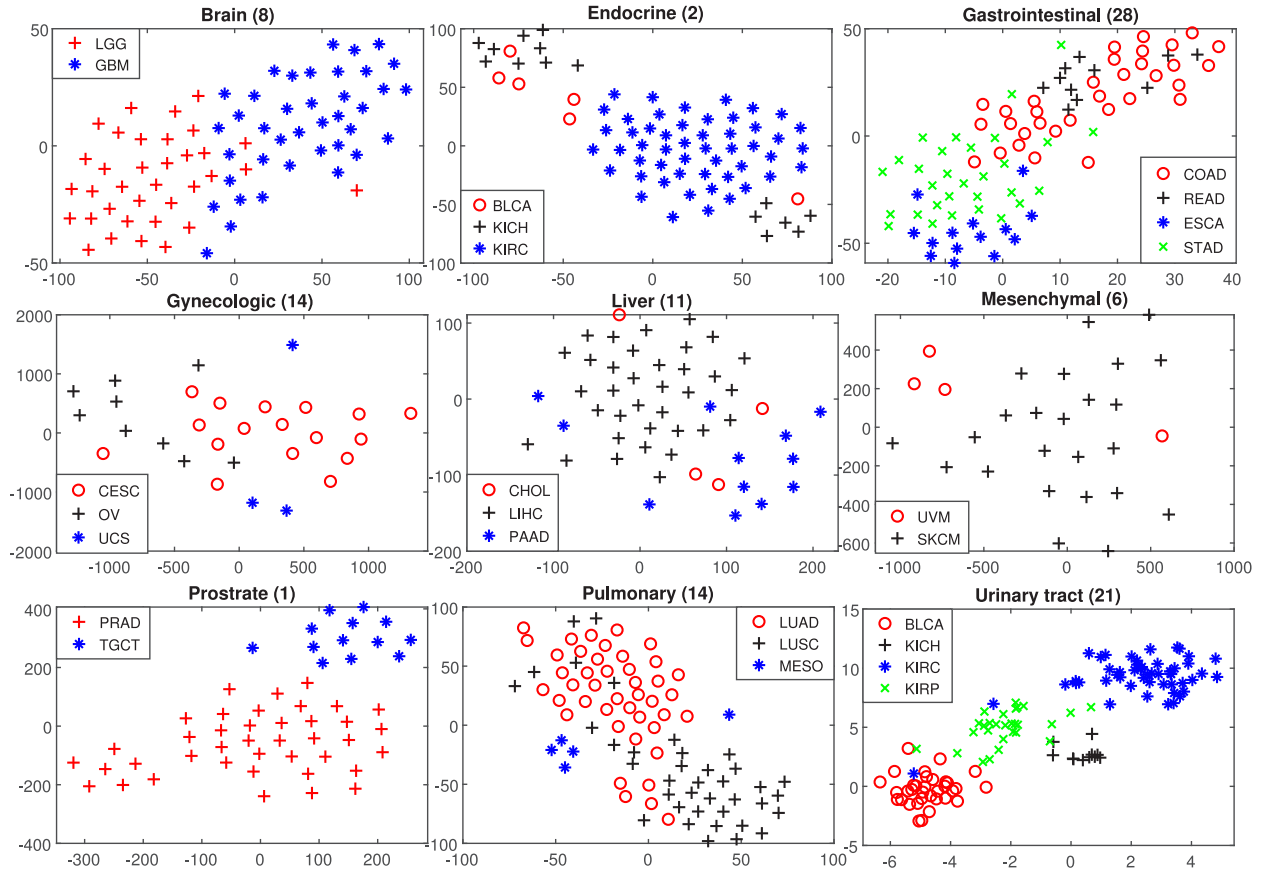


Fig. 2. t-SNE visualization of test set samples using evolutionary features of CFV. For each disease type, a different number of features is selected by the optimizer (presented in Table 2). Mapping the space constructed by optimized features to 2-dimensional space illustrates the discrimination between primary diagnosis for most tumor subtypes.

ROC Analysis — The performance of optimized features are also assessed in terms of Receiver Operating Characteristic (ROC) curves and Area Under Curve (AUC). These results are presented in Fig. 3. The ROC curve of each primary diagnosis is generated by 10 most similar images matched using CFV. The ROC curves and the AUC values support the numerical results. As an example, the consistency between the discrimination of subtypes based upon numerical results and ROC plots can be seen for liver tumor. In this category, CHOL has the least accuracy from the numerical results whereas, according to ROC plots, the same result is obtained. As the curves indicate, in some categories, the average AUC reaches to the maximum value which implies the exact discrimination between classes. In other categories, the values are higher than 0.95.

In addition to ROC curves, the confusion matrices of vertical search resulted by evolutionary features are presented in Fig. 4. The CMs confirm the results presented by other assessment method, i.e., ROC curves and t-SNE visualization. The number of false positive and false negative samples reveal the difficulty of discrimination between some tumor subtypes. Table 7 in the Appendix shows sample queries and image retrieval results by evolutionary features. For each case, the three most similar images are presented for subjective assessment of image search results.

4.3. Evaluation of alternative approaches

In order to compare the proposed evolutionary feature selection method with other commonly-used dimensionality reduction algorithms, Principle Component Analysis (PCA) [52], Autoencoders (AE) [53], and an ANN-based method are utilized to shorten representative vectors. Whereas PCA can provide a few principal features through statistical analysis, Autoencoders compress the features into a small

format. In addition, we used an ANN embedded with L2 regularization method to select the best features according to the resulted coefficients which can show the effect of features. By this method, the features with higher number of very low coefficients (i.e., close to zero) are removed. For the sake of fair comparison, the number of selected features is set to the size of CFV and the threshold to define the prominence of a coefficient is set to 10^{-4} . The regularization rate is set to 0.5 while the number of neurons of fully connected layer is set to $\frac{2}{3} \times$ total number of features. Train, test, and validation sets are similar to those we used for other experiments. A comparison with PCA, Autoencoder, and ANN-based method is provided in Table 3. For each set of mean features extracted from DNN, the methods are applied to map the feature space from 1024 to the number of dimensions with maximum F1-score. For instance, PCA could achieve the maximum F1-score on WSIS from the brain by only two principle components. The same experiments are conducted for Autoencoder so that the neural network is trained with various feature vector lengths. The first layer of encoder part is in the size of all features (i.e., 1024); the size of next layers is achieved by reducing the size of the previous one to its half. For conducting a fair comparison, the best code size from an interval (i.e., (1–50)) is selected to be reported in the table. According to the main objective of this study, PCA, Autoencoder, and NN-based feature selector can be employed as dimension reduction methods, as illustrated in Table 3, where the evolutionary features outperform all alternative methods. As it can be seen, the resulted F1-score using CFV is 93.23% while PCA, Autoencoder, and ANN-based feature selector could achieve 84.83%, 70.11%, and 68.43% of accuracy, respectively. For some cancer types, the ANN-based model cannot distinguish the primary diagnosis accurately. For instance, the resulted F1-score by selected features of ANN-based feature selector is 0% on UVM and MESO cases. Also, on other types, the CVF achieves significantly more

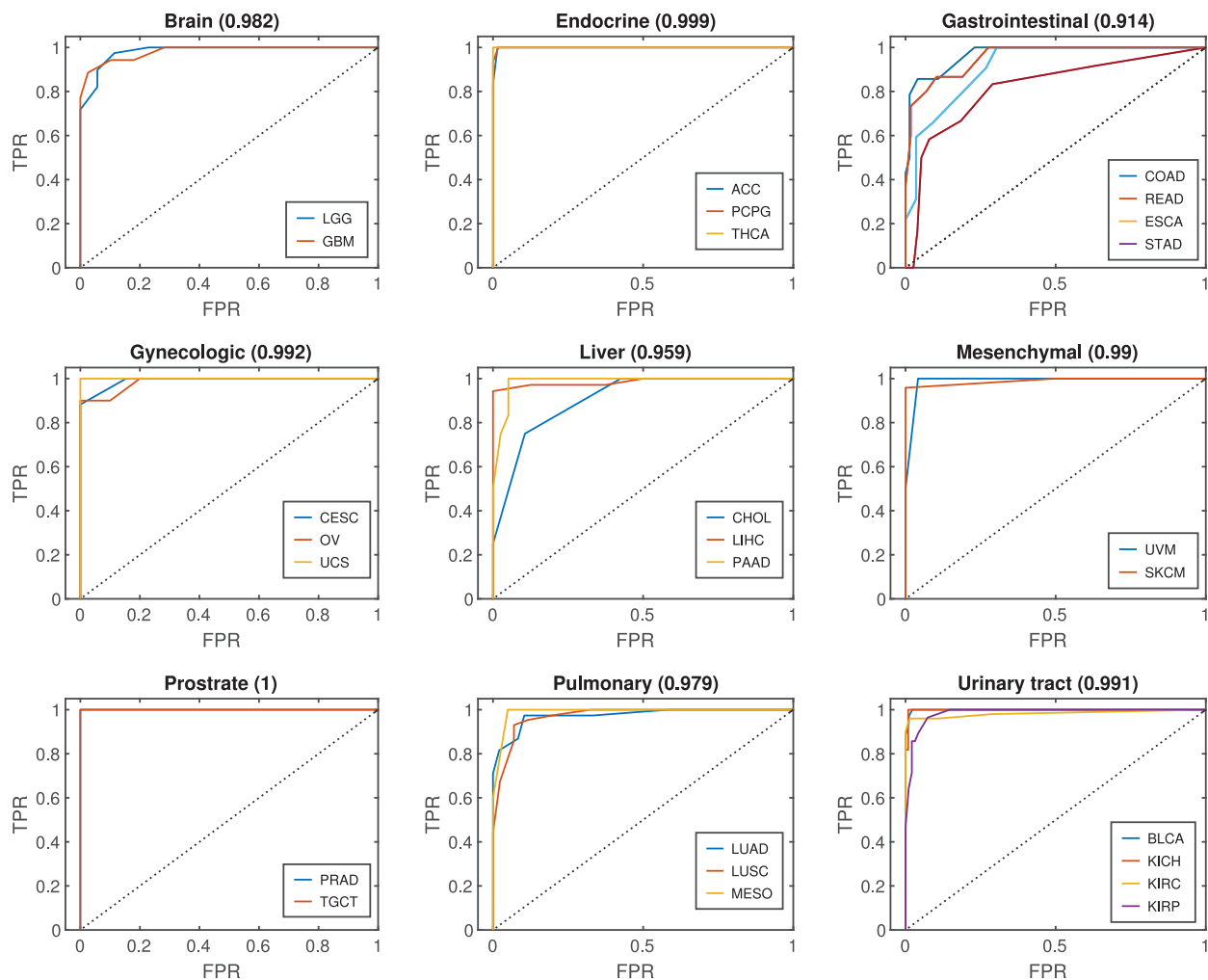


Fig. 3. ROC curves of image retrieval via CFV of each subtype. FPR and TPR stand for false positive rate and true positive rate, respectively. The average Area Under Curve (AUC) over all primary diagnoses is represented at the top of each plot. The minimum average AUC (0.914) is obtained in Gastrointestinal group which results from the difficult distinction between rectal versus colon adenocarcinoma (READ and COAD). That is partly because there can be variability in the morphological appearance of these tumors in general which may be related to degree of differentiation rather than the origin of the tumor i.e. colon vs. rectum. For all other subtypes, the average AUC value is higher than 0.95.

accurate results compared to ANN-based framework. Moreover, these methods require the number of features to be set as a parameter which is a crucial challenge in feature selection domain. Accordingly, they are not the stand-alone methods to select a set of optimal features. For this reason, we required to determine the number of features to be selected by ANN-based approach. Whereas the capability of proposed method to select the best features is an outstanding property. However, even with inspiring this parameter from our proposed method, the competitors are not able to reach the accuracy achieved by evolutionary feature selection.

4.4. Region correspondence matching

In order to verify their expressiveness and generality, we introduce *region correspondence matching* (RCM) to visualize the results related to classification and search. We extracted evolutionary deep features from all tissue patches of each WSI. The pathologist can set – for visualization purposes – the number of tissue types in the WSI, (e.g., three or four). Using organ-specific CFVs, we clustered all WSI patches by the *k*-means algorithm to group tissue patches into four clusters. We observed that the clusters are attributed to histopathologically meaningful areas (see examples in Fig. 5). Fig. 6 shows how RCM can be used to explain image matching. First, deep feature vectors of a query WSI and those of a retrieved WSI are extracted and mixed together. Next, we clustered the

mixed CFVs into three groups using *k*-means clustering and reassigned the patches to query WSI and the top retrieved image. One can now examine the regional correspondence. Fig. 7 shows an example of this similarity explanation scheme. The blue cluster is made of tissue patches of the neoplastic tissue. The orange cluster corresponds to the smooth muscle tissue. The system may direct pathologists to salient features or regions in a tissue specimen that are most important to a given diagnosis. It may also provide an “explainable AI” model to help pathologists gain insight into features most important to a particular diagnosis in the model. This may in turn enhance clinical practice of pathologists in recognizing these features.

5. Summary and conclusions

In this work, we proposed an approach to compress the digital code for representing a WSI. One of the main challenges to design a computer-aided diagnosis system is the large size of the digital histopathology WSIs which can be alleviated by extraction of the most relevant information from those images. Currently, deep learning is a well-known approach that extracts the important information from images. However, the large size of a WSI leads to extraction of many patches with smaller size. The lack of a compressed code for a large WSI is hence a prominent computational challenge which is addressed

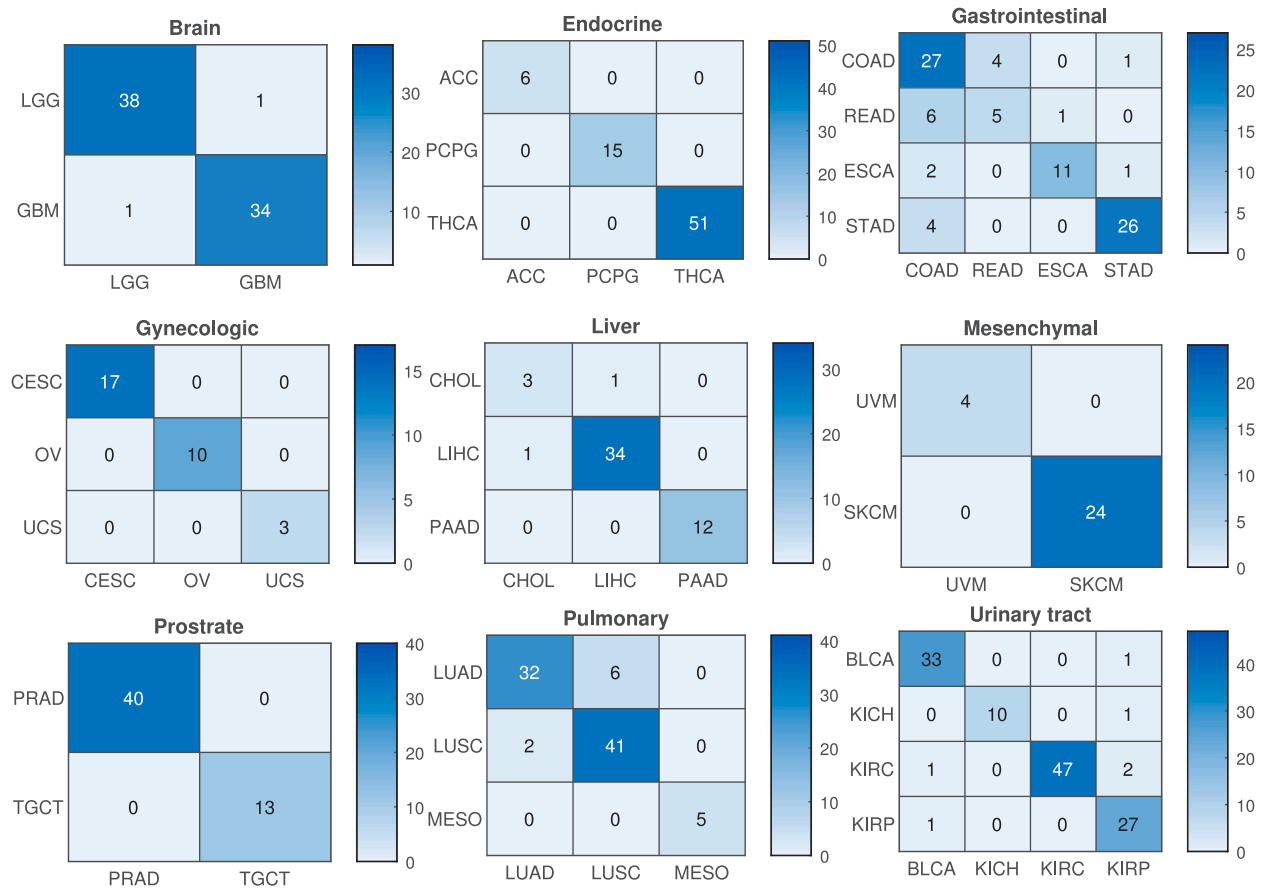


Fig. 4. Confusion matrices of image retrieval on test set of each disease type resulted by evolutionary features. The x-axis indicates the predicted label and y-axis shows the true primary diagnosis. In some cases, the similarity between WSIs causes the false negative or false positive results. For instance, the samples of COAD and READ are predicted incorrectly. Also the STAD and COAD are misclassified in some cases. The same situation occurs between the LUAD and LUAC. Apart from those cases, the results reveal that the evolutionary features exhibit superior class discrimination.

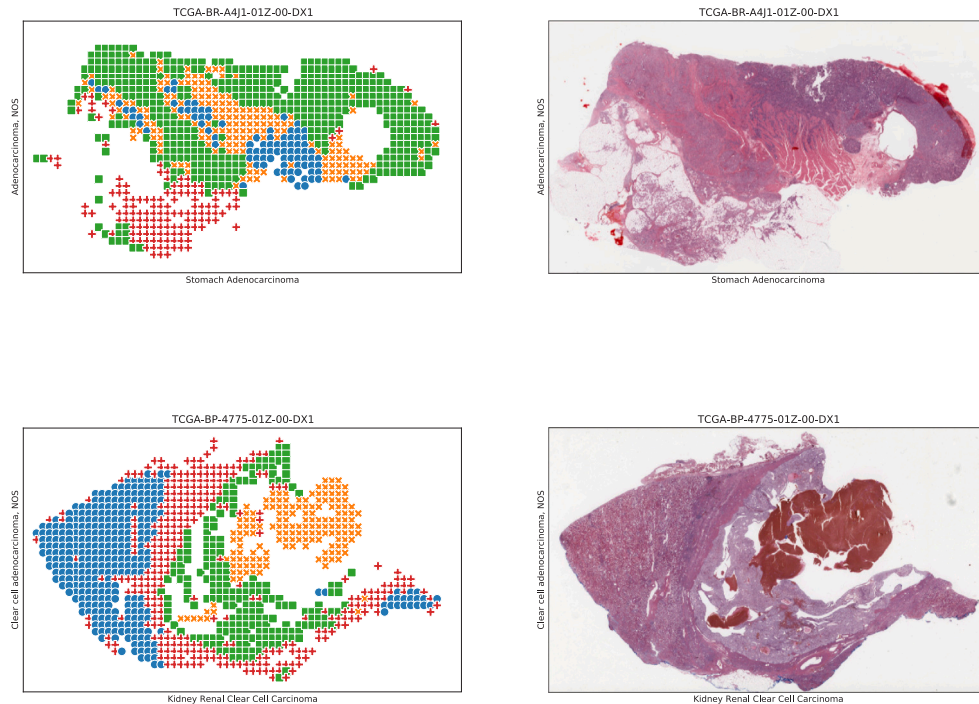


Fig. 5. Deep features were extracted from all 1000 by 1000 pixels patches of two WSIs: Stomach Adenocarcinoma (top), Renal Clear Cell Carcinoma (bottom). For each, organ specific set of features were used for k-means clustering. By gross inspection, it seems that clusters make sense and they divide WSIs into histologically meaningful segments. This finding substantiate the usability of optimal subsets for semantic WSI representation, although tissue patches were represented by deep feature vectors of size smaller than 30.

Table 3

A comparison between the result of the search by proposed method and alternative dimensionality reduction methods PCA, Autoencoders, ANN-based feature selector. The F1-score values of the search of test set resulted from three set of features, evolutionary features, PCA mapped space, autoencoded space, and ANN-based framework are reported. For each method, the number of best subset (i.e., maximum F1-score) is presented. The proposed method could achieve the subset of features which outperforms the alternative methods. The improvement of evolutionary features is around 10%, 22% , and 24.8% higher than PCA, AE, and ANN-based approach, respectively.

Tumor type	Subtype	Proposed CFV		PCA		Autoencoder		ANN-based	
		F1	#F	F1	#F	F1	#F	F1	#F
Brain	LGG	97.43	8	92.75	2	86.2	10	90.66	8
	GBM	97.14		93.67		86.7		90.41	
Endocrine	ACC	100	2	66.66	6	54.54	11	18.18	2
	PCPG	100		86.66		83.87		38.46	
	THCA	100		100		100		87.85	
GI	COAD	76.06	28	68.57	10	68.49	30	69.44	28
	READ	47.62		31.57		11.76		28.57	
	ESCA	84.62		82.75		76.92		60.86	
	STAD	89.66		82.75		80		80	
Gynecologic	CESC	100	14	94.44	16	94.44	29	83.33	14
	OV	100		94.73		85.71		73.68	
	UCS	100		80		0		80	
Liver	CHOL	75	11	66.66	13	33.33	29	75	11
	LIHC	97.14		95.77		97.14		91.66	
	PAAD	100		81.81		84.61		81.81	
Melanocytic	UVM	100	6	66.66	15	75	27	0	6
	SKCM	100		96		95.83		88	
Prostate	PRAD	100	1	100	1	97.56	10	94.6	1
	TGCT	100		100		91.66		85.71	
Pulmonary	LUAD	88.89	14	85.71	3	73.56	30	66.66	14
	LUSC	91.11		83.33		75		71.73	
	MESO	100		88.88		0		0	
Urinary tract	BLCA	95.65	21	92.95	14	80	26	85.71	21
	KICH	95.24		90		43.47		80	
	KIRC	96.91		96		70.96		91.08	
	KIRP	91.53		87.27		76		65.45	
Average		93.23	12	84.83	9	70.11	22	68.43	12

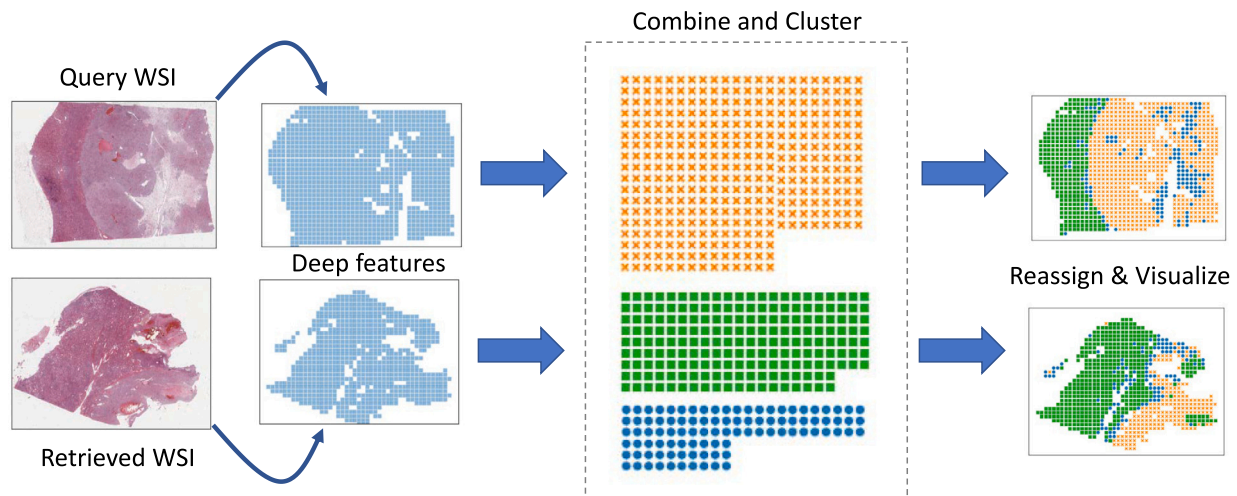


Fig. 6. Region Correspondence Matching (RCM): The proposed feature selection makes storing features for visualization after retrieval. First, compact (short) optimized feature vectors of tissue patches of both query and retrieved WSIs were extracted. Then, features vectors were mixed together, and clustered into an arbitrary number of classes (3 in this case). Tissue patches and their corresponding clusters were reassigned to the WSIs. Consequently, comparable (explainable) visualizations are generated for the retrieval process.

in this study. The developed workflow is proposed to reach two substantial aims, (1) to increase the quality of features encompassing the descriptive visual properties of an image, and (2) to decrease the required data to represent a WSI. These fundamental goals augment digital processing of pathology images such as (1) smaller code provides fast retrieval among millions of images, (2) the visualization and

interpretation of pathology images and image retrieval results are more convenient, (3) using the knowledge obtained from optimal features can lead to designing more compact and efficient deep neural networks, and (4) the demand for large amounts of memory to save descriptive data of biopsy samples decreases markedly. To this end, deep features are extracted from each patch and the mean of feature vectors of all

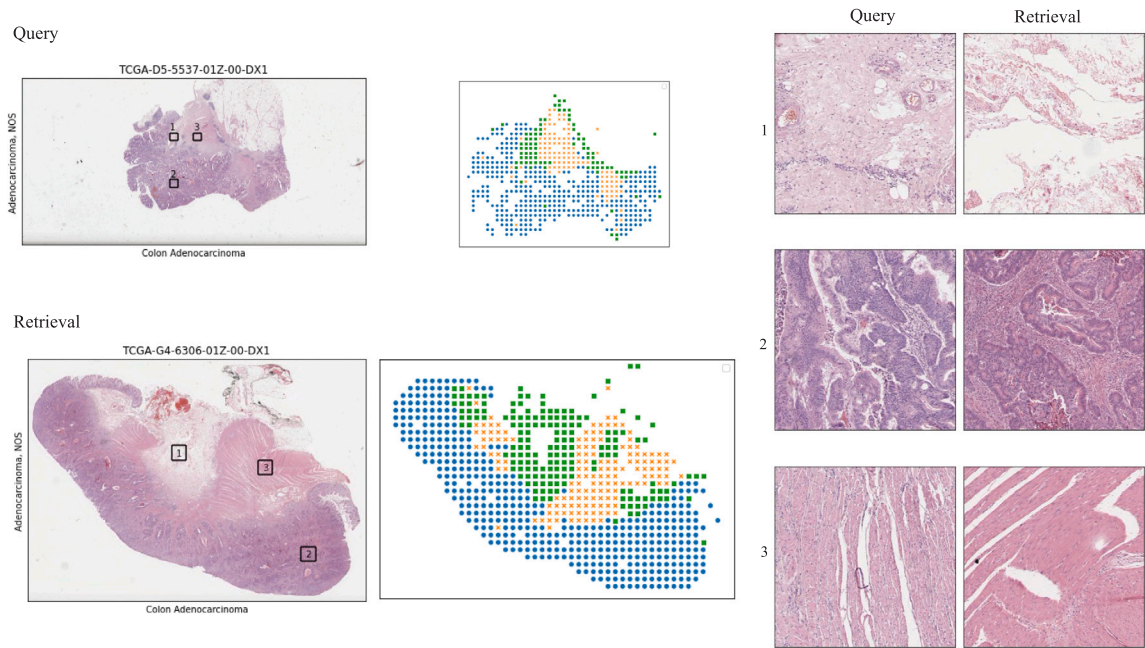


Fig. 7. Example for image matching explanation through CFV clustering. Two digital slides from the TCGA database of colon adenocarcinoma are shown: TCGA-D5-5537-01Z-00-DX1 and TCGA-G4-6306-01Z-00-DX1. Representative image patches are shown from similar regions of these samples including (1) mesenteric attachment, (2), moderately differentiated adenocarcinoma comprised of malignant glands and desmoplastic stroma, and (3) muscularis of the colon wall.

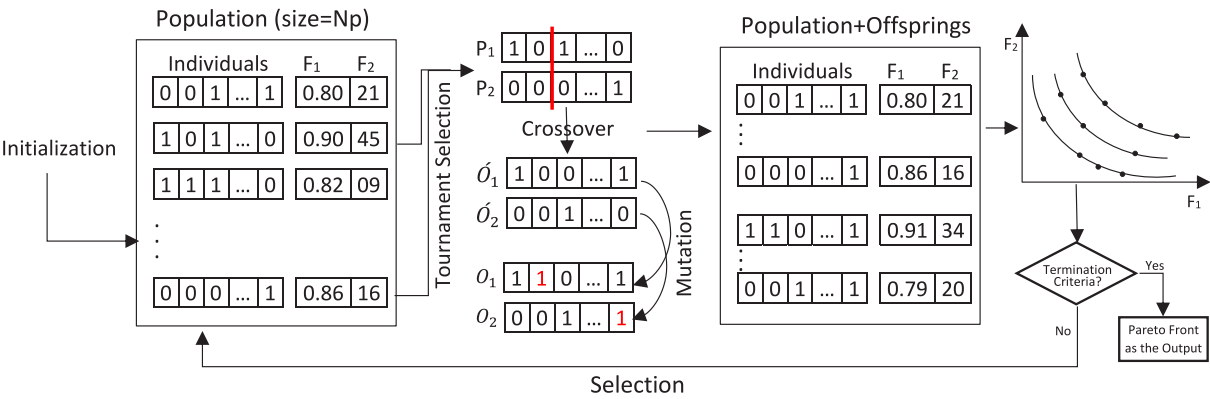


Fig. 8. Evolutionary optimization using binary encoding of the feature vectors. The initial population is generated by random binary vectors which are subsets of feature vectors. F1-score and number of features (as the second objective) are calculated for each individual feature vector. The crossover operator combines any pair of individuals by exchanging their bits. Mutation is applied as well to invert a few bits. The selection technique chooses the best candidate solutions afterwards to update the population for the next generation.

patches corresponding to one WSI provides an efficient representative for such a digital slide. Correspondingly, the evolutionary features selected from mean features not only lead to a compressed code which facilitates histopathology diagnosis workflow, but also augments the accuracy of search or classification of images. The results reveal that if the features are specialized based on tumor type, a more effective image retrieval system can be achieved. The average F1-score for 9 categories of tumor type is 93.23% which indicates 11.96% improvement compared to the original features extracted from the deep neural network. In addition, the average size of the code to represent a biopsy sample decreases to 12 which is 11,093 times smaller than the original features. In addition to numerical results, the quality of

optimal features is indicated by histological visualization. Based upon optimal features, the histological patterns are readily demonstrated. The distinction between different segments in a histopathology image is obviously meaningful. Thus, the visualization of tissue segmentation may assist a pathologist for more efficient screening assessment of image retrieval results. For example, the system may direct pathologists to salient features or regions in a tissue specimen that are most important to a given diagnosis. It may also help pathologists gain insight into features most important to a particular diagnosis in the model. This may in turn enhance clinical practice of pathologists in recognizing these features.

CRediT authorship contribution statement

Azam Asilian Bidgoli: Model designing, Formal analysis, Manuscript drafting. **Shahryar Rahnamayan:** Supervised the method design, Edited the manuscript. **Taher Dehkharghanian:** Manuscript drafting. **Abtin Riasatian:** Manuscript drafting, Preprocessed the images, Designed the deep learning model. **Shivam Kalra:** Preprocessed the images, Designed the deep learning model. **Manit Zaveri:** Preprocessed the images, Designed the deep learning model. **Clinton J.V. Campbell:** Writing – review & editing, Assessed the results. **Anil Parwani:** Writing – review & editing, Assessed the results. **Liron Pantanowitz:** Writing – review & editing, Assessed the results. **H.R. Tizhoosh:** Conceived the study, Supervised the method design, Edited the manuscript.

Declaration of competing interest

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

Acknowledgment

This work is supported by Ontario Research Fund (ORF).

Appendix

See Figs. 8–10 and Tables 4–7.

Table 4

The definition of 32 primary tumor diagnoses and the number of patients in The Cancer Genome Atlas (TCGA) dataset. Each patient might have more than one slide.

Index	ID	Primary diagnosis	#Patients
1	ACC	Adrenocortical Carcinoma	86
2	BLCA	Bladder Urothelial Carcinoma	410
3	BRCA	Breast Invasive Carcinoma	1097
4	CESC	Cervical Squamous Cell Carcinoma and Endocervical Adenoc.	304
5	CHOL	Cholangiocarcinoma	51
6	COAD	Colon Adenocarcinoma	459
7	DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	48
8	ESCA	Esophageal Carcinoma	185
9	GBM	Glioblastoma Multiforme	604
10	HNSC	Head and Neck Squamous Cell Carcinoma	473
11	KICH	Kidney Chromophobe	112
12	KIRC	Kidney Renal Clear Cell Carcinoma	537
13	KIRP	Kidney Renal Papillary Cell Carcinoma	290
14	LGG	Brain Lower Grade Glioma	513
15	LIHC	Liver Hepatocellular Carcinoma	376
16	LUAD	Lung Adenocarcinoma	522
17	LUSC	Lung Squamous Cell Carcinoma	504
18	MESO	Mesothelioma	86
19	OV	Ovarian Serous Cystadenocarcinoma	590
20	PAAD	Pancreatic Adenocarcinoma	185
21	PCPG	Pheochromocytoma and Paraganglioma	179
22	PRAD	Prostate Adenocarcinoma	499
23	READ	Rectum Adenocarcinoma	170
24	SARC	Sarcoma	261
25	SKCM	Skin Cutaneous Melanoma	469
26	STAD	Stomach Adenocarcinoma	442
27	TGCT	Testicular Germ Cell Tumors	150
28	THCA	Thyroid Carcinoma	507
29	THYM	Thymoma	124
30	UCEC	Uterine Corpus Endometrial Carcinoma	558
31	UCS	Uterine Carcinosarcoma	57
32	UVM	Uveal Melanoma	80

Table 5

The results of horizontal search (tumor type identification) by MFVs. Finding the matched WSIs is simply conducted by calculation of Euclidean distance between the mosaics. According to our results, MFVs obtain significantly better search output when compared to the original features. In most cases, search with MFV as one vector of size 1024 achieves higher compared to searching with many feature vectors from many patches. For instance, the melanocytic group of WSIs are retrieved with 96% precision using MFV which is 10% higher than the retrieval result using deep feature vectors of cellMosaic. Therefore, MFV not only compresses the representative vectors of each WSI but also enhances the search performance. Because of the high cellularity of extracted patches, MFV is an efficient representative which encodes the semantic structures of a WSI.

Tumor type	# Patient	All features (130 × 1024)	MFV (1024)	Diff
Brain	74	98.65	98.65	−0.35
Breast	91	91.21	97.80	+6.80
Endocrine	72	91.67	90.28	−1.72
Gastrointestinal	88	84.09	86.36	+2.36
Gynecologic	30	56.67	60	+3
Head/neck	32	87.50	84.38	−3.62
Liver	51	88.24	90.20	+2.19
Melanocytic	28	85.71	96.43	+10.42
Mesenchymal	13	69.23	76.92	+7.92
Prostate/testis	53	96.23	98.11	+2.11
Pulmonary	86	86.05	89.53	+3.53
Urinary tract	123	89.43	91.06	+2.05
Average		85.39	88.31	+2.89

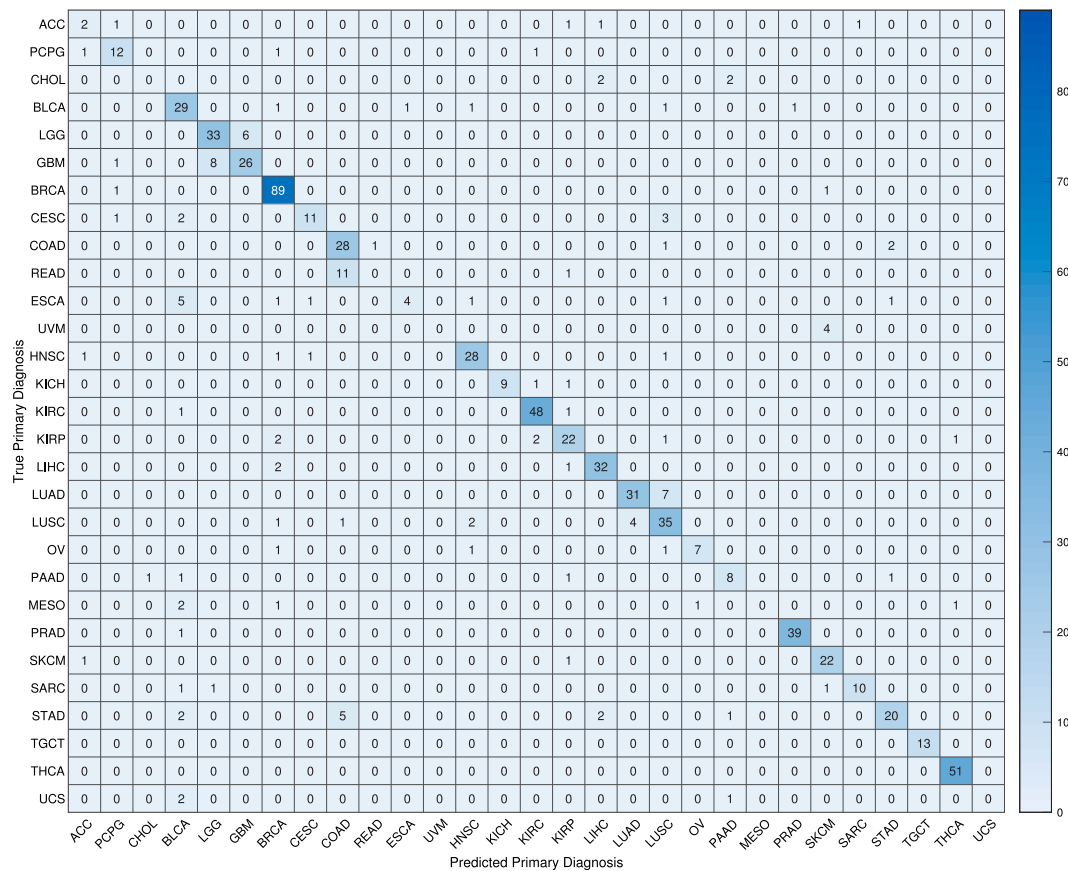


Fig. 9. Confusion Matrix on 29 primary diagnosis by mean features. The test set is searched for each WSI to find the most similar images in terms of primary diagnosis label. The most false negatives and false positives obviously occurred for the subtypes in the same category. For instance, the Rectum Adenocarcinoma (READ) samples are predicted as Colon Adenocarcinoma (COAD) in most of the cases. This was expected because READ and COAD are similar neoplasms of different anatomical sites. These adenocarcinomas are usually morphologically identical. The only way to differentiate them is by anatomic location. This is why most people refer to these groups of cancers collectively as colorectal adenocarcinomas. In addition, there is a significant error to classify the WSIs of MESO, ESCA, and CHOL using the MVF vectors.

Table 6

The results of the tumor type search for all tumor subtypes using the features extracted by a pre-trained network, i.e., DenseNet. Each WSI of the test set is matched against the images of the same tumor type to identify the correct primary diagnosis. Since the network is not trained using the high cellular regions, the MFV is not an reliable representative of a WSI as the resulted F1-score by MFV is lower than the one by original features. However, the feature selection on MFVs could improve the quality of the code so that around 10% improvement has been achieved by optimal features. In addition, a comparison between the result of the search by evolutionary computing (EC) and alternative dimensionality reduction methods PCA and Autoencoders is provided. The F1-score values of the search of test set resulted from three set of features, evolutionary features, PCA mapped space, and autoencoded space are reported. For each method, the number of best subset (i.e., maximum F1-score) is presented. On both networks, EC could achieve the subset of features which outperforms the alternative methods. The improvement of evolutionary features is around 8% and 7% higher than PCA and AE, respectively.

Tumor Type	Subtype	All Features (130 × 1024)	MFV (1024)	Proposed method		PCA		Autoencoder	
				F1	#F	F1	#F	F1	#F
Brain	LGG	7.76	77.78	84.93	42	81.08	5	80.52	15
	GBM	77.10	78.95	85.33		81.08		78.87	
Endocrine	ACC	25	18.18	71.43	6	44.44	3	40	4
	PCPG	57.14	57.14	74.07		62.07		64	
	THCA	94.44	89.52	95.15		92.45		93.58	
Gastrointestinal	COAD	65	53.66	59.74	23	60.98	83	61.11	22
	READ	22.22	22.22	27.27		25		33.33	
	ESCA	50	20	63.64		20		30	
	STAD	62.96	64.29	58.18		72.41		63.33	
Gynecologic	CESC	88.23	83.33	88.24	163	83.33	19	82.35	17
	OV	66.66	70	81.82		70		66.67	
	UCS	75	0	50		0		40	

(continued on next page)

Table 6 (continued).

Tumor Type	Subtype	All Features (130 × 1024)	MFV (1024)	Proposed method		PCA		Autoencoder	
				F1	#F	F1	#F	F1	#F
Liver	CHOL	28.57	0	44.44		50		28.57	
	LIHC	86.15	85.29	91.67	1	89.86	4	85.29	15
	PAAD	69.56	56	85.71		88		66.67	
Melanocytic	UVM	0	40	85.71		66.67		85.71	
	SKCM	92.30	94.12	97.96	50	96	9	97.96	11
Prostrate	PRAD	99	96.39	98.77		96.39		95.12	
	TGCT	96	86.96	96	73	86.96	2	83.33	17
Pulmonary	LUAD	64.93	67.47	69.23		68.24		67.50	
	LUSC	68.88	66.67	71.91	7	65.85	13	71.26	14
	MESO	0	0	0		0		0	
Urinary tract	BLCA	89.85	82.67	81.58		80		75.32	
	KICH	47.61	38.10	50		40		52.63	
	KIRC	82.69	78.79	84.54	62	80	21	74.23	11
	KIRP	76.92	70.59	80.70		70.59		79.25	
Average		63.72	57.62	72.23	47	64.28	18	65.25	14

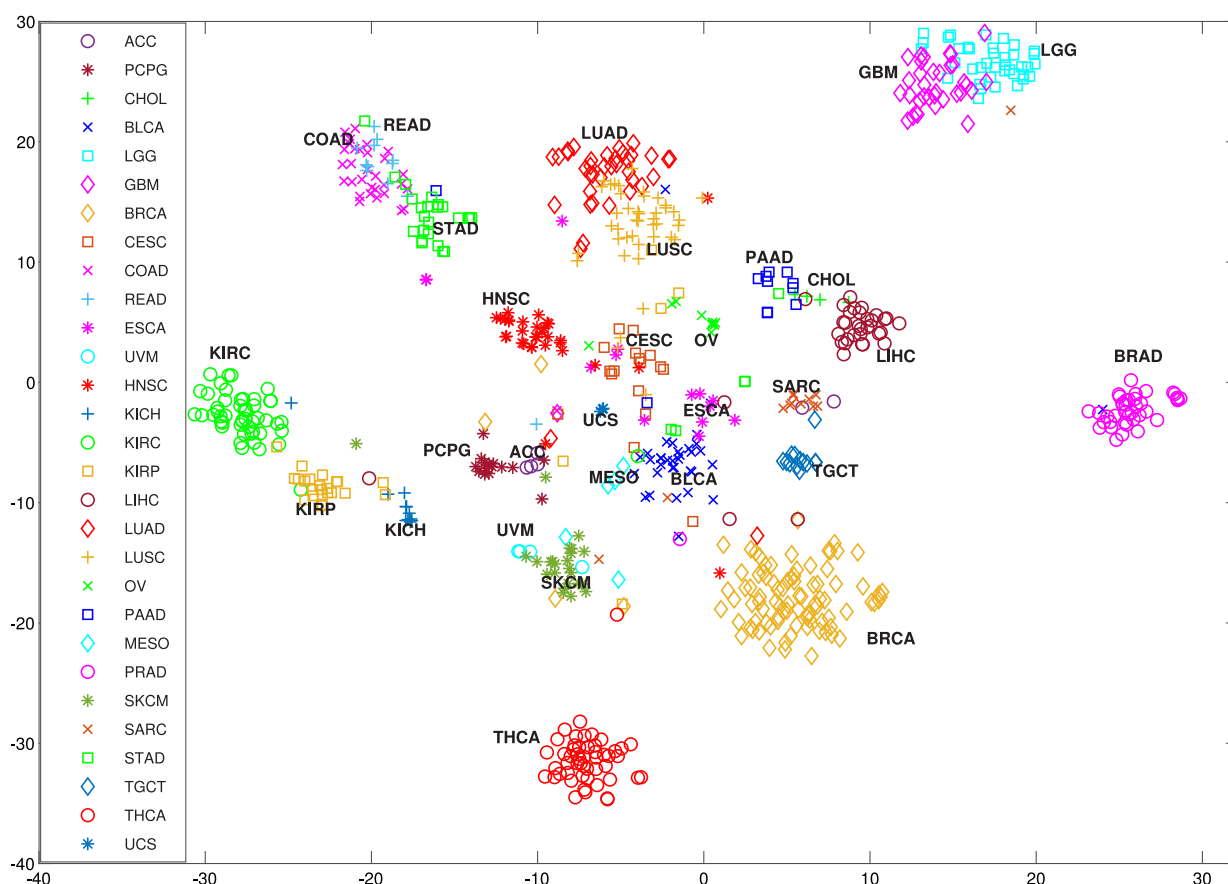
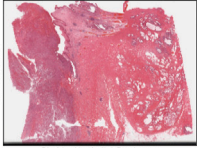
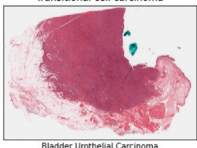
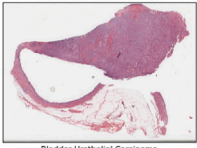
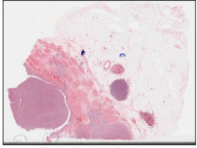
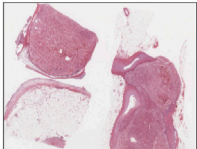
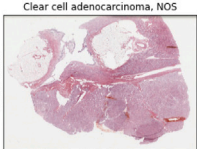
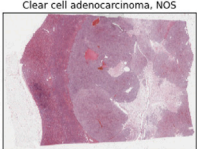
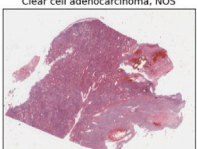
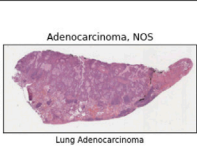
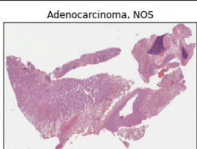
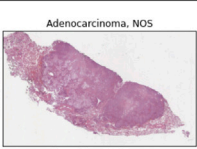
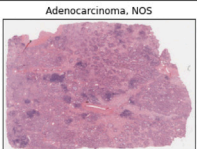
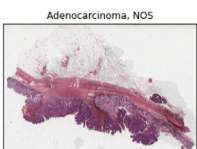
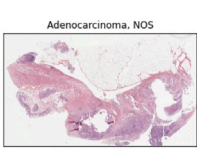
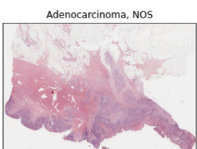
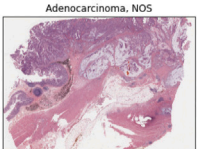

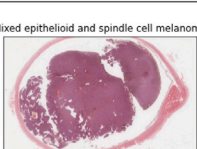
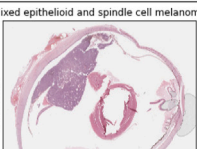
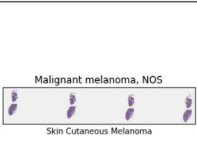


Fig. 10. t-SNE visualization of test set samples using mean features. From the visualization of each category, mean features distinct the primary diagnosis adequately. For most of the groups, the samples are gathered in a dense cluster except for those with a limited number of patients. For instance, according to visualized classes, the Esophageal Carcinoma (ESCA) samples are mixed up with Bladder Urothelial Carcinoma (BLCA) which results in 5 False Negative (FN) samples of ESCA predicted as BLCA according to confusion matrix.

Table 7

Sample results of the vertical search. The left WSI is the query image and the subsequent WSIs are the three most similar images found by matching the CFVs.

Query	First	Second	Third
 Transitional cell carcinoma TCGA-ZF-AA53 Bladder Urothelial Carcinoma	 Transitional cell carcinoma TCGA-ZF-AA54 Bladder Urothelial Carcinoma	 Transitional cell carcinoma TCGA-XF-A9T4 Bladder Urothelial Carcinoma	 Transitional cell carcinoma TCGA-ZF-A9RN Bladder Urothelial Carcinoma
TCGA-ZF-AA53	TCGA-ZF-AA54	TCGA-XF-A9T4	TCGA-ZF-A9RN
 Clear cell adenocarcinoma, NOS TCGA-CZ-5987 Kidney Renal Clear Cell Carcinoma	 Clear cell adenocarcinoma, NOS TCGA-CZ-5451 Kidney Renal Clear Cell Carcinoma	 Clear cell adenocarcinoma, NOS TCGA-B0-5699 Kidney Renal Clear Cell Carcinoma	 Clear cell adenocarcinoma, NOS TCGA-B0-5712 Kidney Renal Clear Cell Carcinoma
TCGA-CZ-5987	TCGA-CZ-5451	TCGA-B0-5699	TCGA-B0-5712
 Adenocarcinoma, NOS TCGA-55-8203 Lung Adenocarcinoma	 Adenocarcinoma, NOS TCGA-55-8514 Lung Adenocarcinoma	 Adenocarcinoma, NOS TCGA-55-8091 Lung Adenocarcinoma	 Adenocarcinoma, NOS TCGA-55-8616 Lung Adenocarcinoma
TCGA-55-8203	TCGA-55-8514	TCGA-55-8091	TCGA-55-8616
 Adenocarcinoma, NOS TCGA-AZ-4682 Colon Adenocarcinoma	 Adenocarcinoma, NOS TCGA-G4-6314 Colon Adenocarcinoma	 Adenocarcinoma, NOS TCGA-G5-6233 Rectum Adenocarcinoma	 Adenocarcinoma, NOS TCGA-CK-5913 Colon Adenocarcinoma
TCGA-AZ-4682	TCGA-G4-6314	TCGA-G5-6233	TCGA-CK-5913
 Mixed epithelioid and spindle cell melanoma TCGA-V4-A9ES Uveal Melanoma	 Mixed epithelioid and spindle cell melanoma TCGA-YZ-A984 Uveal Melanoma	 Mixed epithelioid and spindle cell melanoma TCGA-V4-A9F7 Uveal Melanoma	 Malignant melanoma, NOS TCGA-DA-A1HY Skin Cutaneous Melanoma
TCGA-V4-A9ES	TCGA-YZ-A984	TCGA-V4-A9F7	TCGA-DA-A1HY

References

- [1] Parwani AV. Next generation diagnostic pathology: Use of digital pathology and artificial intelligence tools to augment a pathological diagnosis. *Diagn Pathol* 2019;14(1).
- [2] Schaumberg AJ, Juarez-Nicanor WC, Choudhury SJ, Pastroán LG, Pritt BS, Pozuelo MP, et al. Interpretable multimodal deep learning for real-time pan-tissue pan-disease pathology search on social media. *Mod Pathol* 2020;33(11):2169–85.
- [3] Gutman DA, Cobb J, Somanna D, Park Y, Wang F, Kurc T, et al. Cancer digital slide archive: An informatics resource to support integrated in silico analysis of TCGA pathology data. *J Amer Med Inf Assoc* 2013;20(6):1091–8.
- [4] Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. ImageNet: A Large-scale hierarchical image database. In: *CVPR09*. 2009.
- [5] Zheng Y, Jiang Z, Zhang H, Xie F, Ma Y, Shi H, et al. Histopathological whole slide image analysis using context-based CBIR. *IEEE Trans Med Imaging* 2018;37(7):1641–52.
- [6] Hegde N, Hipp JD, Liu Y, Emmert-Buck M, Reif E, et al. Similar image search for histopathology: SMILY. *Npj Digit Med* 2019;2(1).
- [7] Kalra S, Tizhoosh H, Shah S, Choi C, Damaskinos S, Safarpour A, et al. Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ Digit Med* 2020;3(1):1–15.
- [8] Kurmi Y, Chaurasia V. Content-based image retrieval algorithm for nuclei segmentation in histopathology images. *Multimedia Tools Appl* 2021;80(2):3017–37.
- [9] Hsu W, Markey MK, Wang MD. Biomedical imaging informatics in the era of precision medicine: Progress, challenges, and opportunities. *J Amer Med Inf Assoc* 2013;20(6):1010–3.
- [10] Kumar MD, Babaie M, Tizhoosh HR. Deep barcodes for fast retrieval of histopathology scans. In: *2018 International joint conference on neural networks. IEEE*; 2018, p. 1–8.
- [11] Tizhoosh HR, Zhu S, Lo H, Chaudhari V, Mehdi T. Minmax radon barcodes for medical image retrieval. In: *International symposium on visual computing. Springer*; 2016, p. 617–27.
- [12] Shi X, Sapkota M, Xing F, Liu F, Cui L, Yang L. Pairwise based deep ranking hashing for histopathology image classification and retrieval. *Pattern Recognit* 2018;81:14–22.
- [13] Shi X, Xing F, Xu K, Xie Y, Su H, Yang L. Supervised graph hashing for histopathology image retrieval and classification. *Med Image Anal* 2017;42:117–28.
- [14] Tellez D, Höppener D, Verhoef C, Grünhagen D, Nierop P, Drozdal M, et al. Extending unsupervised neural image compression with supervised multitask learning. In: *Medical imaging with deep learning. PMLR*; 2020, p. 770–83.
- [15] Helin H, Tolonen T, Ylinen O, Tolonen P, Näpänkangas J, Isola J. Optimized JPEG 2000 compression for efficient storage of histopathological whole-slide images. *J Pathol Inform* 2018;9.
- [16] Tellez D, Litjens G, van der Laak J, Ciompi F. Neural image compression for gigapixel histopathology image analysis. *IEEE Trans Pattern Anal Mach Intell* 2019;43(2):567–78.

- [17] Faust K, Bala S, Ommeren RV, Portante A, Qawahmed RA, Djuric U, et al. Intelligent feature engineering and ontological mapping of brain tumour histomorphologies by deep learning. *Nat Mach Intell* 2019;1(7):316–21.
- [18] Pantanowitz L, Farahani N, Parwani A. Whole slide imaging in pathology: Advantages, limitations, and emerging perspectives. *Pathol Lab Med Int* 2015;7:23–33.
- [19] Schaefer R, Otálora S, Jimenez-del Toro O, Atzori M, Müller H. Deep learning-based retrieval system for gigapixel histopathology cases and the open access literature. *J Pathol Inform* 2019;10.
- [20] Komura D, Fukuta K, Tominaga K, Kawabe A, Koda H, Suzuki R, et al. Luigi: Large-scale histopathological image retrieval system using deep texture representations. 2018, *BioRxiv*.
- [21] Kalra S, Tizhoosh H, Choi C, Shah S, Diamandis P, Campbell CJ, et al. Yottixel – An image search engine for large archives of histopathology whole slide images. *Med Image Anal* 2020;65:101757.
- [22] Tomczak K, Czerwińska P, Wiznerowicz M. Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Współczesna Onkol* 2015;1A:68–77.
- [23] Cilia ND, De Stefano C, Fontanella F, di Freca AS. A ranking-based feature selection approach for handwritten character recognition. *Pattern Recognit Lett* 2019;121:77–86.
- [24] Wang C, Hu Q, Wang X, Chen D, Qian Y, Dong Z. Feature selection based on neighborhood discrimination index. *IEEE Trans Neural Netw Learn Syst* 2017;29(7):2986–99.
- [25] Chatterjee R, Maitra T, Islam SH, Hassan MM, Alamri A, Fortino G. A novel machine learning based feature selection for motor imagery EEG signal classification in internet of medical things environment. *Future Gener Comput Syst* 2019;98:419–34.
- [26] Xue B, Zhang M, Browne WN. Particle swarm optimization for feature selection in classification: A multi-objective approach. *IEEE Trans Cybern* 2013;43(6):1656–71.
- [27] Xue B, Zhang M, Browne WN, Yao X. A survey on evolutionary computation approaches to feature selection. *IEEE Trans Evol Comput* 2015;20(4):606–26.
- [28] Hosseini ES, Moattar MH. Evolutionary feature subsets selection based on interaction information for high dimensional imbalanced data classification. *Appl Soft Comput* 2019;82:105581.
- [29] Bidgoli AA, Ebrahimpour-Komleh H, Rahnamayan S. Reference-point-based multi-objective optimization algorithm with opposition-based voting scheme for multi-label feature selection. *Inform Sci* 2021;547:1–17.
- [30] Bar Y, Diamant I, Wolf L, Lieberman S, Konen E, Greenspan H. Chest pathology identification using deep feature selection with non-medical training. *Comput Methods Biomech Biomed Eng: Imaging Vis* 2018;6(3):259–63.
- [31] Toğaçar M, Ergen B, Cömert Z, Özyurt F. A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models. *Irbm* 2020;41(4):212–22.
- [32] Özyurt F. Efficient deep feature selection for remote sensing image recognition with fused deep learning architectures. *J Supercomput* 2019;1–19.
- [33] Mirzaei A, Pourahmadi V, Soltani M, Sheikhzadeh H. Deep feature selection using a teacher-student network. *Neurocomputing* 2020;383:396–408.
- [34] Faust K, Xie Q, Han D, Goyle K, Volynskaya Z, Djuric U, et al. Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. *BMC Bioinformatics* 2018;19(1):1–15.
- [35] Chen Z, Pang M, Zhao Z, Li S, Miao R, Zhang Y, et al. Feature selection may improve deep neural networks for the bioinformatics problems. *Bioinformatics* 2020;36(5):1542–52.
- [36] Riasatian A, Babaie M, Maleki D, Kalra S, Valipour M, Hemati S, et al. Fine-tuning and training of densenet for histopathology image representation using TCGA diagnostic slides. *Med Image Anal* 2021;70:102032.
- [37] Cooper LA, Demicco EG, Saltz JH, Powell RT, Rao A, Lazar AJ. Pancancer insights from the cancer genome atlas: The pathologist's perspective. *J Pathol* 2018;244(5):512–24.
- [38] Boureau Y-L, Ponce J, LeCun Y. A theoretical analysis of feature pooling in visual recognition. In: *Proceedings of the 27th international conference on machine learning*. 2010, p. 111–8.
- [39] Dong X, Yu Z, Cao W, Shi Y, Ma Q. A survey on ensemble learning. *Front Comput Sci* 2020;14(2):241–58.
- [40] Bonawitz K, Eichner H, Grieskamp W, Huba D, Ingerman A, Ivanov V, et al. Towards federated learning at scale: System design. 2019, *arXiv preprint arXiv:1902.01046*.
- [41] Al-Tashi Q, Abdulkadir SJ, Rais HM, Mirjalili S, Alhussian H. Approaches to multi-objective feature selection: A systematic literature review. *IEEE Access* 2020;8:125076–96.
- [42] Dy JG, Brodley CE, Kak A, Broderick LS, Aisen AM. Unsupervised feature selection applied to content-based retrieval of lung images. *IEEE Trans Pattern Anal Mach Intell* 2003;25(3):373–8.
- [43] Kharrat A, Mahmoud N. Feature selection based on hybrid optimization for magnetic resonance imaging brain tumor classification and segmentation. *Appl Med Inf* 2019;41(1):9–23.
- [44] Li A-D, Xue B, Zhang M. Multi-objective feature selection using hybridization of a genetic algorithm and direct multisearch for key quality characteristic selection. *Inform Sci* 2020.
- [45] Kimeswenger S, Tschandl P, Noack P, Hofmarcher M, Rumetshofer E, Kindermann H, et al. Artificial neural networks and pathologists recognize basal cell carcinomas based on different histological patterns. *Mod Pathol* 2020;1–9.
- [46] Chan A, Tuszynski JA. Automatic prediction of tumour malignancy in breast cancer with fractal dimension. *R Soc Open Sci* 2016;3(12):160558.
- [47] de Paula LC, Soares AS, de Lima TW, Coelho CJ. Feature selection using genetic algorithm: An analysis of the bias-property for one-point crossover. In: *Proceedings of the 2016 on genetic and evolutionary computation conference companion*. 2016, p. 1461–2.
- [48] Lim SM, Sultan ABM, Sulaiman MN, Mustapha A, Leong KY. Crossover and mutation operators of genetic algorithms. *Int J Mach Learn Comput* 2017;7(1):9–12.
- [49] Deb K, Jain H. An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: Solving problems with box constraints. *IEEE Trans Evol Comput* 2013;18(4):577–601.
- [50] Tahir MA, Bouridane A, Kurugollu F. Simultaneous feature selection and feature weighting using hybrid Tabu search/K-nearest neighbor classifier. *Pattern Recognit Lett* 2007;28(4):438–46.
- [51] Demšar J. Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 2006;7:1–30.
- [52] Ringnér M. What is principal component analysis? *Nature Biotechnol* 2008;26(3):303–4.
- [53] Yang Y, Wu QJ, Wang Y. Autoencoder with invertible functions for dimension reduction and image reconstruction. *IEEE Trans Syst, Man, Cybern: Syst* 2016;48(7):1065–79.