Journal Pre-proof

Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopatological images

Patrik Sabol, Peter Sinčák, Pitoyo Hartono, Pavel Kočan,
Zuzana Benetinová, Alžbeta Blichárová, Ľudmila Verbóová,
Erika Štammová, Antónia Sabolová-Fabianová, Anna Jašková

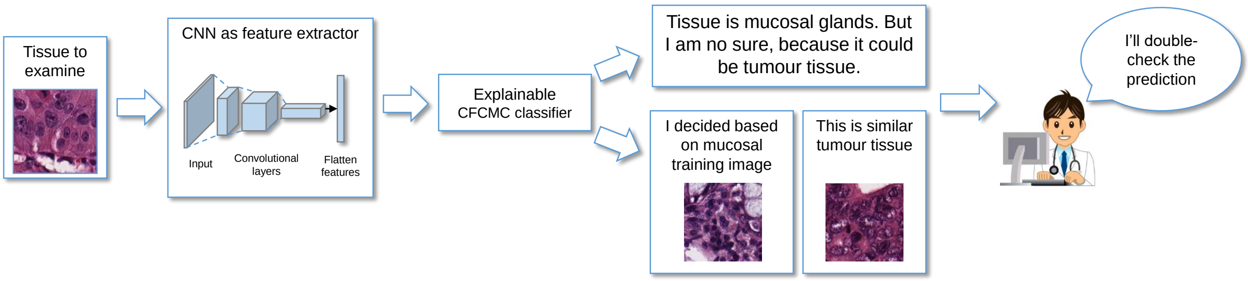
PII: S1532-0464(20)30151-9
DOI: https://doi.org/10.1016/j.jbi.2020.103523
Reference: YJBIN 103523

To appear in: *Journal of Biomedical Informatics*

Received date : 12 April 2020
Revised date : 22 July 2020
Accepted date : 27 July 2020

Please cite this article as: P. Sabol, P. Sinčák, P. Hartono et al., Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopatological images, *Journal of Biomedical Informatics* (2020), doi: https://doi.org/10.1016/j.jbi.2020.103523.

*Graphical Abstract


Tissue to examine
CNN as feature extractor
Input
Convolutional layers
Flatten features
Explainable CFCMC classifier
Tissue is mucosal glands. But I am no sure, because it could be tumour tissue.
I decided based on mucosal training image
This is similar tumour tissue
I'll double-check the prediction

*Highlights (for review)

- An explainable CFCMC classifier used for classification of eight tissue type from histopathological colorectal cancer images
- Its classification performance was significantly enhanced using a fine-tuned Convolutional Neural Network
- Explanation interface for segmentation of the whole-slide images that provides a semantical and visual explanation was developed
- The system was designed to be used as a support system for medical experts
- The results from clinical trials with 14 pathologists indicate the usefulness and reliability of the proposed classifier and its explanations

Explainable Classifier for Improving the Accountability in Decision-Making for Colorectal Cancer Diagnosis from Histopatological Images

Patrik Sabol, Peter Sinčák*

Department of Cybernetics and Artificial Intelligence, Technical University of Košice, Košice, Slovakia

Pitoyo Hartono

School of Engineering, Chukyo University, Nagoya, Japan

Pavel Kočan, Zuzana Benetinová, Alžbeta Blichárová, Ľudmila Verbóová, Erika Štammová

Department of Pathology, Pavol Jozef Šafárik University in Košice, Košice, Slovakia

Antónia Sabolová-Fabianová, Anna Jašková

The Faculty of Arts of Prešov University, Prešov, Slovakia

Abstract

Pathologists are responsible for cancer type diagnoses from histopathological cancer tissues. However, it is known that microscopic examination is tedious and time-consuming. In recent years, a long list of machine learning approaches to image classification and whole-slide segmentation has been developed to support pathologists. Although many showed exceptional performances, the majority of them are not able to rationalize their decisions. In this study, we developed an explainable classifier to support decision making for medical diagnoses. The proposed model does not provide an explanation about the causality between the input and the decisions, but offers a human-friendly explanation about the plausibility of the decision. Cumulative Fuzzy Class Membership Criterion (CFCMC) explains its decisions in three ways: through a semantical explanation about the possibilities of misclassification, showing the training sample responsible for a certain prediction and showing training samples from conflicting classes. In this paper, we explain about the mathematical structure of the classifier, which is not designed to be used as a fully automated diagnosis tool but as a support system for medical experts. We also report on the accuracy of the classifier against real world histopathological data for colorectal

*Corresponding author
Email address: patrik.sabol@tuke.sk (Patrik Sabol, Peter Sinčák)

cancer. We also tested the acceptability of the system through clinical trials by 14 pathologists. We show that the proposed classifier is comparable to state of the art neural networks in accuracy, but more importantly it is more acceptable to be used by human experts as a diagnosis tool in the medical domain.

Keywords: Explainable artificial intelligence, explainable machine learning, uncertainty measure, digital pathology, colorectal cancer

1. Introduction

Histopathological images of cancer tissue are routinely analysed by pathologists who are responsible for the cancer type diagnosis and prognosis [1]. Different types of tissues can be distinguished from histopathological evaluations of Hematoxylin and Eosin (H&E) stained tissue sections. In colorectal cancer (CRC), the tumour architecture changes during tumour progression and is related to patient prognosis [2]. Microscopic examination of tissue sections is known to be tedious and time-consuming [3]. Furthermore, the outcome of the analysis may be affected by the levels of experience of the pathologists involved. Therefore, with the advancement of digital pathology, computer-aided analysis of the histopathological images and machine learning-based diagnostic systems, the fidelity and efficacy of medical diagnoses can be significantly improved.

A long list of machine learning approaches to image classification and whole-slide segmentation has been developed to support pathologist in interpreting histopathological images [4, 5]. Especially, in the recent years, conventional classification approaches, which mainly rely on manually-engineered features, were outperformed by Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) [6]. The high performances of deep models is due to their ability to automatically extract representations that are strongly relevant to the their predictions from the learning data. However, their performance is not supported by their ability to explain their decisions and hence, may prevent their applicability in real world clinical settings [7]. Our objective in this study is to build a reliable classifier that is also able to provide explanations about its decision in human-friendly forms. We believe that in real world clinical settings, which require accountability, the accuracy of a classifier should be paired with its explainability [8].

1.1. Relevant studies

In [9], *Holzinger et al.* distinguished two types of explainable AI: *Ante-hoc systems* which incorporate explainability directly into the structure of an AI-model; these are systems that are interpretable by design. Typical examples include linear regression, decision trees and fuzzy inference systems. They are commonly referred to as *white-boxes* or, currently, *glass-boxes* [10]. *Posthoc systems*, on the other hand, aim to explain and interpret *black-box* classifiers which provide local explanations for their specific decision. The majority of the explanation approaches seek to link a particular output of the classifier to input variables to see the impact of features on the final decision outcome. For instance, in [11], *G. R. Vásquez-Morales et al.* used neural network-based classifier to predict whether a person is at risk of developing chronic kidney disease. Here, a black-box machine-learning method was complemented by Case-Based Reasoning, a white-box method that is able to find explanatory cases for an explanation-by-example justification of a neural network's prediction. In [12], *Mullenbach et al.* presented an attentional convolutional network that predicted medical codes from clinical texts. Using an attention mechanism, the most relevant segments of the clinical text for each of the medical codes were selected and used as an explanation mechanism. Through an interpretability evaluation by a physician, they showed that the attention mechanism identified meaningful explanations. In [13], *Lundberg et al.* presented an ensemble-model-based machine learning method using deep learning that predicts the near-term risk of hypoxaemia during anaesthesia care and explains the patient and surgery-specific factors that led to that risk. The system improved the performance of anesthesiologists by providing interpretable hypoxaemia risk and the contributing factors. In [7], *Hagele et al.* utilized Layer-wise Relevance Propagation (LRP) to provide pixel-level explanation heatmaps for the classification decision of the CNN in digital histopathology analyses of tumour tissue. These explanations were used to improve the generalization of the classifier by detecting and removing the effects of hidden biases in used datasets. A similar approach to visualize parts of the input image responsible for the prediction was used in [14], where LIME (Local Interpretable Model-agnostic Explanations) was utilized to provide a global understanding for the CNN model by providing explanations for individual instances in the context of in-vivo gastral image analysis.

It is natural that in the delicate medical domain, prediction models should not only be accurate, but also accountable; they should state uncertainty in their predictions, indicating difficult cases for which further human expert inspections are necessary. Therefore, another approach to the probe and interpret machine learning algorithm is to measure the uncertainty of the prediction for one particular example, the *predictive uncertainty* [9]. In [15], a transparent neural network, S-rRBF, was proposed and applied to DNA microarray data sets. It provides an intuitive explanation through a visualization of its decision process and on the given problem. It allows the users to understand why a certain problem is easy or difficult. Moreover, it makes it possible to see whether a new input is hard to classify or unlikely to be misclassified. However, the visual information still needs to be interpreted and thus is prone to subjective inconsistencies. For the field of digital pathology, in [16], *Raczkowski et al.* proposed an accurate, reliable and active (ARA) image classification framework using a Bayesian Convolutional Neural Network (ARA-CNN) for classifying histopathological images of colorectal cancer. The model is able to achieve reliability by measuring the uncertainty of each prediction. This capability was used to identify mislabelled training samples. In [17], the recently proposed semantically explainable fuzzy classifier called Cumulative Fuzzy Class Membership Criterion (CFCMC) [18, 19] was used to classify histopathology images for breast cancer and to generate additional information about classification reliability in human-friendly terms, in the form of a semantic explanation. It provides a confidence measure for the classification result of a test image followed by a visualization of training image and the most similar images that belong to clusters of the conflicting class with a different confidence degree. In this paper, we extend the explainability of the CFCMC classifier by defining the *factor of miclassification (FoM)* and the *certainty threshold*. While the *FoM* is a value that describes the possibility of the input sample being misclassified to the one particular conflicting class, *the certainty threshold* is a value of the *FoM*, under which it is a certain that the input sample will not be misclassified. Compared to the concept of the uncertainty measure proposed in [16], in the case of uncertain prediction, our approach is additionally able to suggest the classes in which the input sample could be misclassified. Thus, it offers relevant classes to be further examined.

Different approach to interpret the decision of the classifier is based on generating instances that are close to an observation. In [20] the influence function was used to trace a model's prediction through the learning algorithm and back to its training data, thereby identifying training points most responsible for a given prediction. This approach was used to explain the prediction of a blackbox, deep neural network model. Moreover, the paper [21] investigates the effects of presentation of influence of training data points on machine learning predictions to boost user trust by measuring psychological signals (Galvanic Skin Response and Blood Volume Pulse). It showed that these features correlate to user trust. Such reference-based explanations are needed in medicine, where, for instance, they could help to diagnose the type of the cancer from histopathological images.

In this study the explainability of our classifier refers to its ability in providing a degree of confidence for each of its prediction and in expressing the information in intuitive and human-friendly manners. The information is expressed through visualizations of training examples that are responsible for the prediction outcome paired with sematical explanations regarding likelihood for misclassifications. Here, while our method does not provide explainability for the causality between the input and the prediction, we believe that the explainability improves the accountability of the proposed classifier, and thus significantly contributes in supporting decision making in time-crucial medical domains. The objective of this article is to apply the explainable CFCMC classifier for the classification of histopathological images of colorectal cancer. We used a publicly available dataset that was released in [2] by *Kather et al.*. It consists of a training set comprised of 5000 small tiles, each of them annotated with one of eight tissue classes and 10 non-annotated whole slide images (WSI) of the tissue.

In [16], it was shown that CNN outperformed the approach in [2], where features derived from images using texture descriptors served as a basis for a support vector machine model to classify colorectal cancer. Moreover, in [22], CNN achieved an exceptional level of performance, 98,7% accuracy, in nine tissue types classifications of colorectal cancer, using the VGG19 model [23], which was pretrained on the ImageNet database [24]. Therefore, to enhance the accuracy of the CFCMC, we employ a Convolutional Neural Network as a feature extractor. We are aware of the problem of losing explainablity of the CNN model by compressing of the data from the feature space to the latent space, which causes that it is hard to track the decision back to the features in the feature space. This problem is not relevant for our study, because we do not provide an explanation about the causality between the features and the decisions but we provide explanation about the classifiability of the data, which is significantly improved by the CNN model by mapping the data from the feature space to the latent space.

Finally, we developed an explanation interface, which provided a semantical and visual explanation that was extracted from the CFCMC classifier that was used to classify the WSIs of the colorectal cancer tissue. We evaluated our XAI (eXplainable Artificial Inteligence) system using common within-subject experimental design [25]; the outcomes from our explanation interface (XAI system) were compared with outcomes from a stand alone CNN (AI system with no explanation) by 14 pathologists at clinical trials through questionnaires.

2. Proposed explainable model

In this section, the mathematical description of the Cumulative Fuzzy Class Membership Criterion classifier is explained and followed by the definition of the factor of misclassification and the certainty threshold.

2.1. Cumulative Fuzzy Class Membership Criterion decision based classifier

The proposed method is based on the assumption that d-dimensional data in the feature space are split into n_c classes, where C_i $(i = 1, \cdots, n_c)$, the i-th class, is divided into n_{cl}^i clusters, where Cl_{ij} $(j = 1, \cdots, n_{cl}^i)$ is the j-th cluster of the i-th class. Each cluster Cl_{ij} comprises training data $\widetilde{p}_{ijk} \in \mathbb{R}^d, (k = 1, \cdots, m_{ij})$, and m_{ij} is the number of training patterns of cluster C_{ij}. Each training pattern $\widetilde{p}$ defines a fuzzy class membership criterion $\kappa_{\widetilde{p}}(\overline{x})$, which is considered as a triangular function as follows:

$$\kappa_{\widetilde{p}_{ijk}}(\overline{x}) = \begin{cases} 1 - \dfrac{\|\widetilde{p}_{ijk} - \overline{x}\|}{a_{ij}} & \|\widetilde{p}_{ijk} - \overline{x}\| < a_{ij} \\ 0 & otherwise \end{cases}, \tag{1}$$

where, $\overline{x} \in \mathbb{R}^d$ is an input vector, a_{ij} is the width of the triangular function for the j-th cluster of i-th class.

Let $(\kappa_{\widetilde{p}_{ij1}}, \ldots, \kappa_{\widetilde{p}_{ijK_{ij}}}, \ldots, \kappa_{\widetilde{p}_{ijm_{ij}}})$ be a desceding order of the fuzzy class membership criterion $\kappa_{\widetilde{p}_{ij}}$ for the j-th cluster of i-th class such that $\kappa_{\widetilde{p}_{ij1}} \geq \cdots \geq \kappa_{\widetilde{p}_{ijm_{ij}}}$, where K_{ij} is the number of first values in the reordering.

Then, the Cumulative Fuzzy Class Membership Criterion (CFCMC) for class C_i is defined as follows:

$$\chi_{C_i}\left(\overline{x}\right) = \max_j \left(\tfrac{1}{K_{ij}} \sum_{k=1}^{K_{ij}} \kappa_{\widetilde{p}_{ijk}}\left(\overline{x}\right) \right), \tag{2}$$

where, $\chi_{C_i}\left(\overline{x}\right)$ is the value of CFCMC for an unknown pattern $\overline{x}$ to the class C_i.

Then, the decision rule for winner class CL for the input pattern $\overline{x}$ is as follows:

$$CL(\overline{x}) = C_{\underset{i}{\operatorname{argmax}}\left(\chi_{C_i}\left(\overline{x}\right)\right)}. \tag{3}$$

2.2. Algorithm description

The algorithm consists of two phases: the initialization and the learning phase. The initialization phase consists of three processes: data splitting, clustering, and parameters initialization. First, input data are divided into three sets: training sets, validation sets, and testing sets. Training patterns are used in Eq. 1 to create a CFCMC decision surface. During the learning phase, the decision surface is optimized (parameters a_{ij} Eq. 1 and $K_{i,j}$ in Eq. 2 are adaptively optimized) in order to cover all validation patterns. The testing set is used for the final evaluation of the created decision surface.

Afterwards, the training data of each class C_i are independently clustered in order to find n_{cl}^i clusters for each class in feature space using the well-known K-means algorithm. The number of clusters, k, is estimated via a gap statistic [26]. This technique uses the output of any clustering algorithm, comparing the change of the within-cluster dispersion with that which is expected under an appropriate reference null distribution. Any other techniques for the estimation of the number of clusters can be used, such as Silhouette analysis [27] or Davies-Bouldin clustering criterion [28].

Next, the parameters a_{ij} and K_{ij} in Eq. 1 and in Eq. 2, respectively, are initialized. These parameters affect the shape of the boundary created by fuzzy class membership criterion $\kappa_{\widetilde{p}}$. Every fuzzy class membership criterion $\kappa_{\widetilde{p}_{ijk}}$ of the j-th cluster of the i-th class shares the same value of parameters a and K. a_{init} is initialized as follows:

$$a_{init} = \tfrac{1}{n_{\widetilde{p}}} \sum_{ij=1}^{n_{\widetilde{p}}} \min_{j \neq i} \|\widetilde{p}_i - \widetilde{p}_j\|, \tag{4}$$

where, $n_{\widetilde{p}}$ is the number of training patterns. K_{init} value is initialized from the interval $(1; m_{ij})$. The value of threshold θ is set from the interval $(0; 1)$. If the value of CFCMC $\chi(\overline{x})$ of the input pattern $\overline{x}$ is below the threshold θ, thus $\chi(\overline{x}) < \theta$, the pattern is "not classified". Finally, the CFCMC surface is computed using Eq. 1 and Eq. 2.

During the learning phase, adjustment of the CFCMC surface's shape occurs in order to obtain the highest classification accuracy. An assumption of dividing training set into n_c classes and each class C_i into n^i_{cl} clusters $Cl_{i,j}$ generates a set of vectors

$$\overline{p}_i = [a_{i1}, K_{i1}; \cdots ; a_{ij}, k_{ij}; \cdots ; a_{in^i_{cl}}, k_{in^i_{cl}}], \tag{5}$$

where, $i = 1, \cdots , n_c$. Optimizing of $\overline{p}_i$, the CFCMC surface's shape is adjusted. Any optimizing algorithm can be used, for instance, simulated annealing or hill-climbing methods. We decided to employ a well-known genetic algorithm [29]. The fitness function is defined as follows:

$$Minimize : err\ (S_{valid})\,, \tag{6}$$

where, S_{valid} is the validation set and $err\,(S)$ is the error rate evaluated from the data set S as follows:

$$err(S) = \frac{\#\text{of incorrectly classified samples}}{\#\text{of all samples}}. \tag{7}$$

2.3. Factor of misclassification

The term factor of misclassification *(FoM)* is described as *"the likehood of the input sample, which is assigned to the cluster Cl belonging to the class C, to be misclassified to one of the rest of the classes"* i.e. the possibility that in reality the observation belongs to another class. The factor of misclassification of the input sample, assigned to the cluster Cl_{Ai}, to the reference cluster Cl_{Bj} is defined as follows:

$$FoM(\overline{x}, Cl_{Bj}) = \frac{\chi_{Cl_{Bj}}(\overline{x})}{\chi_{Cl_{Ai}}(\overline{x})} + sim_{Cl}(Cl_{Ai}, Cl_{Bj}), \tag{8}$$

where the first term on thee right hand side describes the *local similarity* as the ratio between memberships of the input sample $\chi(\overline{x})$ to the reference cluster Cl_{Bj} and the winner cluster Cl_{Ai}. The second term on the left hand side describes *global similarity*, which is based on the relationship between the data's clusters.

The similarity between the two clusters Cl_{Ai} and Cl_{Bj} is defined as follows:

$$sim_{Cl}(Cl_{Ai}, Cl_{Bj}) = \frac{A_{intersection(Cl_{Ai}, Cl_{Bj})}}{A_{Cl_{Ai}}}, \tag{9}$$

where A_{Cl} is the area of a hypersphere describing cluster Cl and $A_{intersection}$ is the area of intersection of the two clusters. Here, for simplification, the clusters are described with n-dimensional hypersphere, where n is the data dimensionality. For the center and the radius of a hypersphere, the coordinates of a cluster's centroid $c_{Cl_{ij}}$ and the estimated variance $\hat{\sigma}_{Cl_{ij}}$ of a cluster's data, respectively, are used. For computational purposes, n-hypersphere is transferred into a two dimensional circle. The area of intersection between the two clusters is computed using simple two dimensional trigonometry by using the distance between the centers and radiuses of the circles, and thus, the Euclidean distance

$\left\| c_{Cl_{Ai}} - c_{Cl_{Bj}} \right\|$ between the centroids of the clusters Cl_{Ai} and Cl_{Bj} and their estimated variances $\hat{\sigma}_{Cl_{Ai}}$ and $\hat{\sigma}_{Cl_{Bj}}$.

The equation for the estimation of the cluster's variance value $\hat{\sigma}_{Cl_{A,i}}$ was derived in [19] and it is calculated as follows:

$$\begin{aligned} \hat{\sigma}_{Cl_{ij}} = a_{Cl_{ij}} \left(\frac{k}{\chi_{Cl_{ij}^{max}}} \right)^{m} &\begin{cases} m = 0.7 & \chi_{Cl_{ij}^{max}} \leq k \\ m = 2.5 & \chi_{Cl_{ij}^{max}} > k \end{cases}, \\ k &= p_1 * K + p_2 \\ p_l &= a_l * dim^{b_l} + c_l \;\; (l = 1, 2) \\ a_1 = -0.7621, b_1 &= -0.2799, c_1 = 0.0746 \\ a_2 = 0.8372, b_2 &= -0.3729, c_2 = 0.1758, \end{aligned} \tag{10}$$

where dim is the dimensionality of the data.

It should be noted that during the variance estimation, the Euclidean distance was replaced with the following distance measure d: Let's have two vectors x_A and x_B with n-dimensionality. Then the distance d is defined as follows:

$$d = \frac{1}{n} \sum_{i=1}^{n} |x_{Ai} - x_{Bi}|, \tag{11}$$

The value of the factor of misclassification of the input $\overline{x}$ to the i-th class C_i is computed as follows:

$$FoM(\overline{x}, C_i) = \max_j FoM(\overline{x}; Cl_{ij}) \tag{12}$$

The factor of misclassification can also be expressed semantically. It exhibits the values as follows:

$$DFoM \begin{cases} no \quad possibility & FoM(\overline{x}, C_A) \in (0, c_\theta) \\ low \quad possibility & FoM(\overline{x}, C_A) \in (c_\theta, \theta_{mid}] \\ high \quad possibility & FoM(\overline{x}, C_A) \in (\theta_{mid}, FoM_{max}] \end{cases} \tag{13}$$

where c_θ is the certainty threshold and $\theta_{mid} = (FoM_{max} - c_\theta)/2 + c_\theta$. The FoM_{max} is the maximum value of the FoM computed using the validation samples as follows:

$$FoM_{max} = \max_j \max_i FoM(\overline{x}_j, C_i), \tag{14}$$

where $\overline{x} \in S_{valid}$.

2.4. Certainty threshold and certain prediction

The certainty threshold is the value of the FoM, below which it is certain (i.e. there is no possibility) that the input sample $\overline{x}$ is assigned to the class C_A and will not be misclassified to any other classes in the feature space.

Let $\overline{x}$ be samples from the validation set S_{valid} that were misclassified and C_{GT}^{k} be the ground truth label of the k-th sample $\overline{x}_k$. Then the certainty threshold c_θ is calculated as follows:

$$c_\theta = \min_{k} FoM(\overline{x}_k; C_{GT}^{k}). \tag{15}$$

It follows that if $FoM(\overline{x}, C_i) < c_{FoM}$ holds, it is unlikely that input sample $\overline{x}$ belongs to the class C_i.

Therefore, if it holds that $\forall i_{\in 1,\cdots,n_c}; FoM(\overline{x}, C_i) < c_\theta$, the prediction of the input sample $\overline{x}$ is *certain*, otherwise it is *uncertain*.

2.5. Representative training sample

Representative training sample $\widetilde{p}^{r}$ is the training sample that is the most responsible for assigning input sample $\overline{x}$ to the j-th cluster Cl_{ij} of the i-th class C_i. It is computed as follows:

$$\widetilde{p}^{r}(\overline{x}) = \operatorname*{argmin}_{k} \|\widetilde{p}_{ijk} - \overline{x}\| \tag{16}$$

where $\|\widetilde{p}_{ijk} - \overline{x}\|$ is the Euclidean distance between $\overline{x}$ and $\widetilde{p}_{ijk}$.

3. Colorectal cancer detection explanation interface

This section describes the explanations generated to human experts by the proposed system in colorectal cancer detection tasks. To evaluate the usefulness of the proposed explanations, we designed two systems; first a plain CNN model that only generates decisions without any explanation, second, an X-CFCMC (eXplainable CFCMC) model that complements its decisions with explanations.

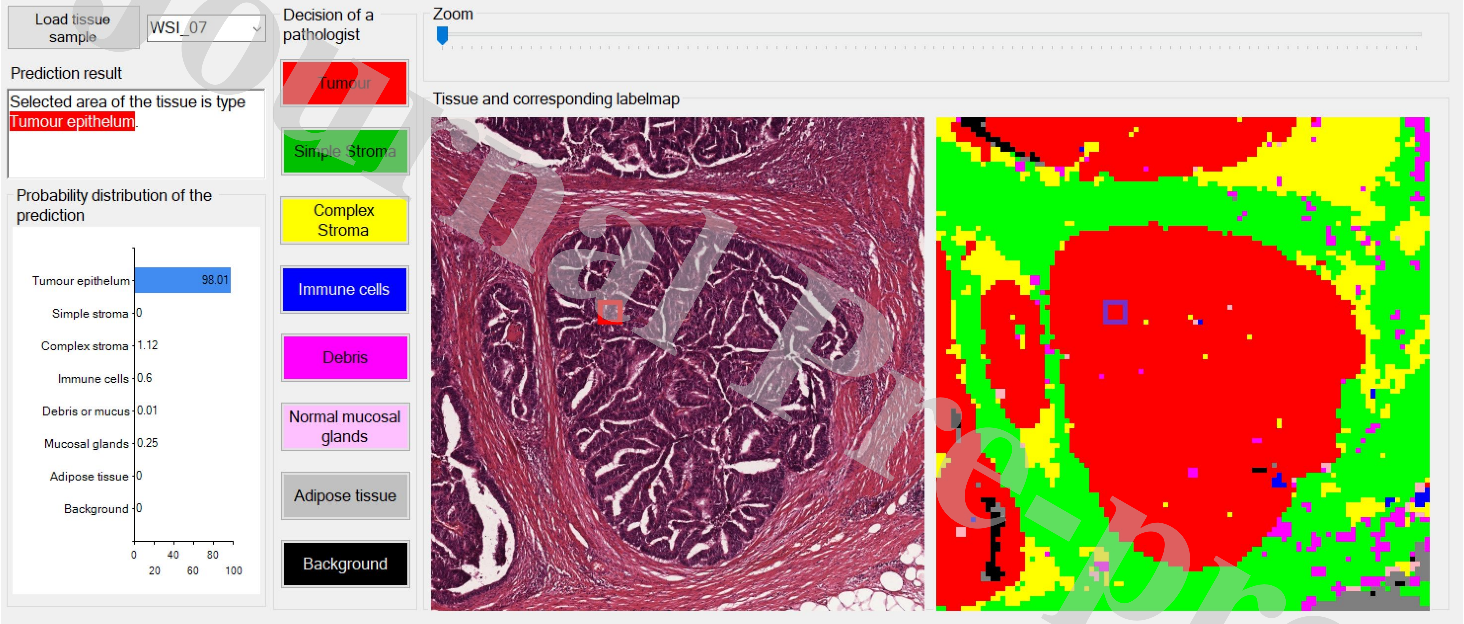
For both systems, we developed similar user interfaces; both provide classification results of the whole-slide images (WSI) of colorectal cancer tissue, showing the original image of the WSI and a corresponding label map with a colour code of eight different tissue types. Pathologist can examine an arbitrary area of the WSI by clicking on the desired area. Subsequently, the interfaces show their predictions. Finally, a pathologist can provide the final decision by selecting one of the eight buttons representing the eight different tissue types.

The *non-explainable AI system* interface that uses the plain CNN model is shown in Fig. 1. It provides predicted type of tissue and a probability distribution of the prediction of all tissue types, which was computed in the output layer of the CNN model with the softmax activation function.

The *explainable AI system* interface that uses the X-CFCMC model is shown in Fig. 2. It complements its decision with three types of information:

3.0.1. Semantical explanation

To provide user-friendly explanations, prediction results and information regarding the possibility of the misclassification of the examined area of the WSI are semantically explained, for example, with the phrase ”there is a low possibility that this classification is wrong”. This semantical explanation is generated based on the value of the FoM.



Selected area of the tissue is type Tumour epithelium.

Figure 1: The interface for the plain CNN, without any explanation, for presenting prediction on histopathological WSI. From the left hand side, it shows a prediction result and the probability distribution of the prediction. Next, eight buttons for making a final decision by a pathologist are located there. Finally, the original image of the WSI and the corresponding label map are visualized.

3.0.2. Visualization of the training image most responsible for a given prediction

To justify the prediction result, the most responsible training image for a given prediction is displayed to the user. A similar approach to understand the predictions was introduced in [20], where the influence function was used to trace the CNN's prediction through the learning algorithm and back to its training data. In our approach, the most responsible training image for a prediction is a representative training sample $\widetilde{p}^r(\overline{x}_w)$ for the input image $\overline{x}_w$ assigned to the winner tissue type w. It follows that if the training image has very similar context with the input image, it should gain the trust of the pathologist in prediction. Otherwise, the pathologist could consider the certain prediction as not being reliable.

3.0.3. Visualization of training images of other types of tissue

The third means of explanation shows the representative training images for the input image to the other tissue types, into which the input image could be misclassified with a high or low possibility. It should visually explain to the pathologist why the input sample could be misclassified to a particular tissue type. In the case of similar context, a pathologist could consider the particular tissue type as the true type of tissue.

4. Experiments

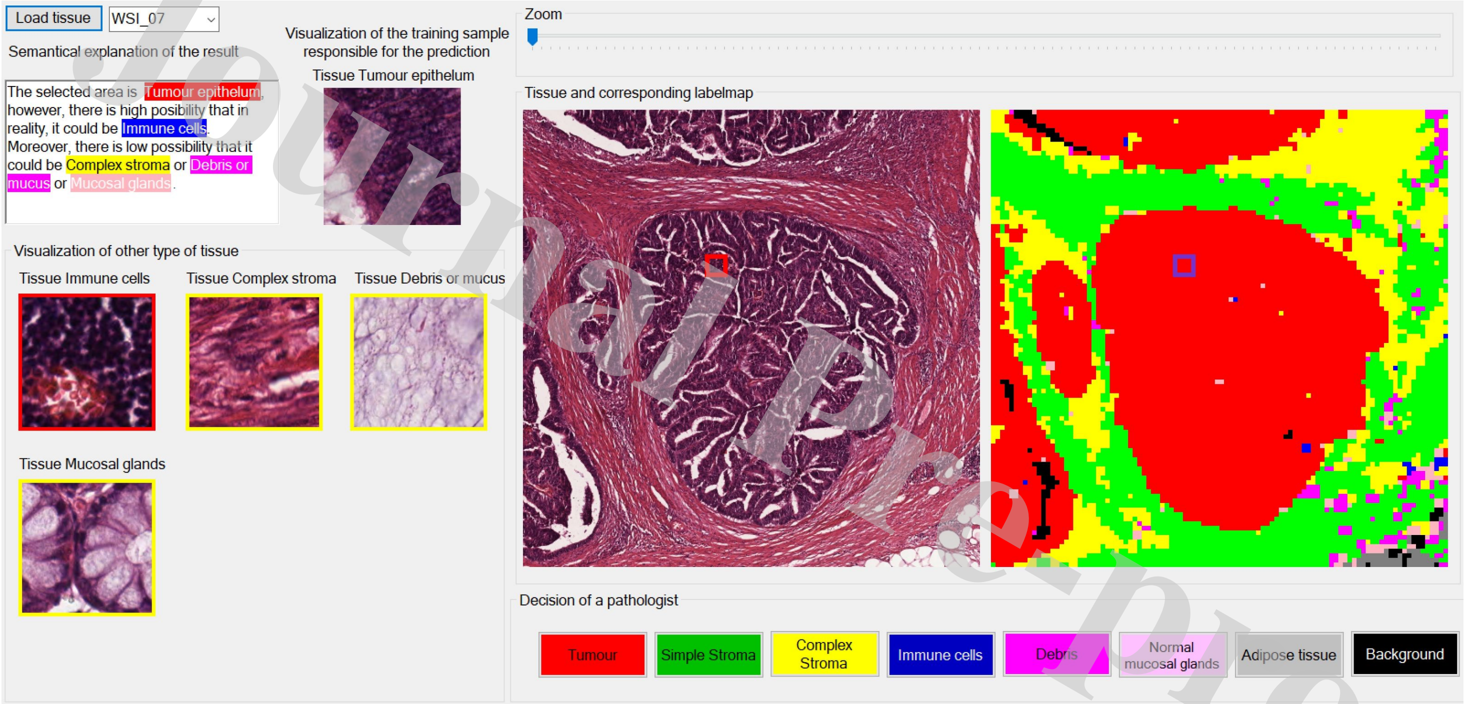
To choose the best performing explainable model to classify colorectal cancer image data, the model has to be accurate and reliable with its explanations. Therefore, in this section, first, we present the results from the task of boosting the performance of the CFCMC classifier. Then, we describe the results of validating the *certainty threshold*. Finally, we show examples of generated explanations.

4.1. Histopathological data description

We used a publicly available dataset released in [2] by *Kather et al.*. It consists of Hematoxylin and Eosin (H&E) tissue slides, which were cut into 5000 small tiles of the size 150x150 pixels (equivalent to 74 μm 74 μm), each of them annotated with one of eight tissue classes, namely tumour epithelium, simple stroma (homogeneous composition, includes tumour stroma, extra-tumoural stroma and smooth muscle), complex stroma (containing single tumour cells and/or few immune cells), immune cells (including immune-cell conglomerates and sub-mucosal lymphoid follicles), debris (including necrosis, haemorrhage and mucus), normal mucosal glands, adipose tissue, background (no tissue). The data are class-balanced, each of the classes consists of 625 tiles.

4.2. Boosting the CFCMC's performance

The focus of the first part of experiments was to find the architecture of the CNN with the best performance as a feature extractor to train the explainable CFCMC classifier.



The selected area is Tumour epithelum, however, there is high posibility that in reality, it could be Immune cells. Moreover, there is low possibility that it could be Complex stroma or Debris or mucus or Mucosal glands.

Figure 2: The interface for the explainable system (X-CFCMC system) for presenting predictions with additional explanations of the CFCMC. From the left hand side, it shows a semantical explanation of the results alongside the visualization of the training sample responsible for the prediction, below which there is a visualization of the other tissue types, which could potentially be the true tissue type. Next, the original image of the WSI and the corresponding label map are visualized, below which eight buttons for making a final decision by a pathologist are located.

4.2.1. Experimental setup

We utilized eight well-known CNN models, pre-trained on ImageNet [24] dataset, specifically, *AlexNet* [30], *VGG-16* [23], *Inception-v3* [31], *ResNet-50* [32], *Xception* [33], *DenseNet121* [34], *Inception-ResNet-V2* [35] and *EfficientNet0* [36]. For all of them, the fully connected layers were cut and replaced with a dense layer containing 1024 neurons with ReLU activation functions and with an output layer containing 8 neurons with the softmax activation function.

Moreover, we created three lighter architectures that were trained from scratch, specifically a *VGG-like* model with 12 convolutional layers and 2 fully-connected layers, an *Inception-like* model with 3 inception layers and a *ResNet20* model with a depth 20. All models were trained using the Adam [37] optimizer to minimize cross-entropy for 100 epochs with the learning rate set to the value 0.0001.

Finally, to train the explainable CFCMC classifier, we extracted the features from the last dense layer of all the CNN models. The experimental setup for the CFCMC algorithm is as follows: number of clusters for each of the classes was set to 1. The value of the threshold θ was set to value $\theta = 0.01$. For the optimization of the CFCMC, MATLAB implementation of the genetic algorithm was used with a population size of 50 individuals. The mutation rate was set to 0.2. Arithmetic crossover and adaptive feasible mutation operators were used for reproduction. Stochastic uniform selection was used to choose parents for the next generation. The algorithm stops at the 30th generation.

4.2.2. Experimental results

Table 1 provides the performance results of different CNN models and the corresponding CFCMC models. The classification results are evaluated with a 10-folds cross validation test. Because of the class-balanced dataset, the accuracy metric was chosen to evaluate the performance.

From Table 1, it can be observed that pre-trained and fine-tuned models outperform the ones trained from scratch. Moreover, it can be seen that features extracted from the CNN models significantly boost the performance of the explainable CFCMC models.

4.3. Validation of the certainty threshold

To validate the certainty threshold, three metrics were defined: certainty rate, certainty error and ground truth label certainty error.

The *certainty rate* c_r is defined as the ratio between the number of certain predictions y_c^* and the number of all predictions y^*.

$$c_r = \frac{\#y_c^*}{\#y^*} \tag{17}$$

The *certainty error* is defined as follows:

$$c_e = \frac{\#\widetilde{y}_c^*}{\#y_c^*} \tag{18}$$

	CNN model	**CFCMC model**
Raw image	—	59.35(3.43)
AlexNet	91.43(2.42)	85.32(3.26)
VGG-16	93.61(1.75)	92.42(2.02)
Inception-v3	92.76(1.44)	90.79(2.04)
ResNet-50	**93.80(1.08)**	91.28(1.64)
Xception	93.58(1.25)	**92.78(1.74)**
DenseNet121	92.76(1.29)	92.06(1.75)
Inception-ResNetV2	92.76(1.02)	91.44(1.95)
EfficientNet0	90.97(1.39)	85.84(5.83)
VGG-like	80.88(1.14)	80.21(2.14)
Inception-like	85.25(2.82)	83.97(3.04)
ResNet20	90.14(1.24)	84.01(5.02)

Table 1: The performance results of different CNN models and the corresponding CFCMC models

where $\widetilde{y}^*_c$ are certain predictions that were misclassified.

The *ground truth label certainty error* c_e^{GT} is defined as follows:

$$c_e^{GT} = \frac{\#\widetilde{y}^*_{GT_c}}{\#\widetilde{y}^*} \tag{19}$$

where $\widetilde{y}^*$ are misclassified predicitions and $\widetilde{y}^*_{GT_c}$ are misclassified predictions, which for their ground truth label C_{GT} holds that $\phi(\overline{x}; C_{GT}) < c_\theta$, i.e. their ground truth label is unlikely to be true label.

Table 2 provides the results from the three previously defined metrics computed on predictions from joined testing sets from 10 folds for each CFCMC model. For each metric, the best performing models are highlighted in bold.

CFCMC models trained on features that were extracted from pre-trained CNN models generally outperformed models trained from scratch with a certainty rate c_r (13.99% against 3.41% in average). All of the models, however, achieved very low certainty error c_e; 7 models achieved zero error, while the highest value was 2.17%. It follows that when a classifier labels its prediction as *certain*, it is unlikely that this prediction will be incorrect.

Moreover, all of the models likewise reached a low value of *ground truth label certainty error* c_e^{GT} (lower than 5%). This implies that when a classifier labels its prediction as *uncertain*, it is very likely that ground truth label will appear among the potentially true labels.

4.4. Performance on imbalanced data set

For the clinical trials we used a balanced dataset, in which each class is uniformly distributed. Typically, this is not the case in many real world conditions; in the most cases, classifiers are required to

CFCMC models	$\#y^*$	$\#\widetilde{y}^*$	err	$\#\ y_c^*$	c_r	$\#\widetilde{y}_c^*$	c_e	$\#\widetilde{y}_{GT_c}^*$	c_e^{GT}
AlexNet	5000	734	*14.68%*	538	*10.76%*	**0**	***0.00%***	6	*0.82%*
VGG16	5000	379	*7.58%*	646	*12.92%*	1	*0.15%*	5	*1.32%*
Inception-v3	5000	460	*9.20%*	513	*10.26%*	**0**	***0.00%***	7	*1.52%*
ResNet-50	5000	436	*8.72%*	589	*11.78%*	**0**	***0.00%***	13	*2.98%*
Xception	5000	**361**	***7.22%***	**1368**	***27.36%***	2	*0.20%*	15	*4.16%*
DenseNet121	5000	397	*7.94%*	648	*12.96%*	**0**	***0.00%***	**4**	**1.01%**
Inception-ResNetV2	5000	428	*8.56%*	612	*12.24%*	**0**	***0.00%***	8	1.87%
EfficientNet0	5000	707	*14.14%*	680	*13.60%*	3	*0.44%*	**4**	**0.57%**
VGG-like	5000	990	*19.80%*	46	*0.92%*	1	*2.17%*	7	0.71%
Inception-like	5000	801	*16.02%*	291	*5.82%*	**0**	***0.00%***	6	0.75%
ResNet20	5000	800	*16.00%*	175	*3.50%*	**0**	***0.00%***	**4**	**0.50%**

Table 2: The results from three metrics for validation of the certainty threshold, certainty rate c_r, certainty error c_e, ground truth label error c_e^{GT} with corresponding occurrences

Class	Training set		Validation set		Testing set	
	Balanced	*Imbalanced*	*Balanced*	*Imbalanced*	*Balanced*	*Imbalanced*
Tumour	0.75	0.30	0.10	0.10	0.15	0.60
Simple stroma	0.75	0.20	0.10	0.10	0.15	0.70
Complex stroma	0.75	0.70	0.10	0.10	0.15	0.20
Immune cells	0.75	0.60	0.10	0.10	0.15	0.30
Debris	0.75	0.70	0.10	0.10	0.15	0.20
Mucosal glands	0.75	0.80	0.10	0.10	0.15	0.10
Adipose	0.75	0.50	0.10	0.10	0.15	0.40
Background	0.75	0.40	0.10	0.10	0.15	0.50

Table 3: Fraction ratios of balanced and imbalanced generated data sets

Metric	CNN		CFCMC	
	Balanced	*Unbalanced*	*Balanced*	*Unbalanced*
Accuracy	92.74%	90.06%	91.08%	87.23%
Precision	92.50%	89.93%	91.44%	90.33%
Recall	92.76%	90.20%	91.04%	89.24%
F1	92.64%	90.08%	91.26%	89.78%

Table 4: The performance results of CNN model with Xception architecture and corresponding CFCMC models for balanced and imbalanced data sets. The results of the precision, recall and F1 metrics are averages of all classes.

Data set	$\#y^*$	$\#\widetilde{y}^*$	err	$\#\ y_c^*$	c_r	$\#\widetilde{y}_c^*$	c_e	$\#\widetilde{y}^*_{GT_c}$	c_e^{GT}
Balanced	750	67	8.74%	113	15.06%	0	0%	2	2.98 %
Unbalanced	1872	243	12.76%	289	15.43%	0	0%	8	3.29 %

Table 5: The results from three metrics for validation of the certainty threshold, certainty rate c_r, certainty error c_e, ground truth label error c_e^{GT} with corresponding occurrences for both balanced and unbalanced data set

deal with imbalances. Therefore, to examine the performance of our classifier on imbalanced data set, we generated two data sets from histopathological images: *balanced data set* with the same ratio and *imbalanced data set* with different ratio for each of the class (See Tab. 3).

Table 4 provides the performance results of CNN models with fine-tuned Xception architecture and corresponding CFCMC models for both data sets. Because we deal with imbalanced data set, beside the accuracy, the classification results are also evaluated with other metrics (precision, recall and F1). The result shows that the differences of the CNN and CFCMC models between balanced and imbalanced data set are not significant, hence the models are able to deal with imbalanced data set.

Moreover, Table 5 presents the results from the metrics for val-

idation of the certainty threshold (defined in 4.3) for balanced and unbalanced data sets. As can be seen, the difference in the results of all metrics are insignificant, which indicates that the FoM is unaffected by imbalance in data set.

4.5. Explanations examples

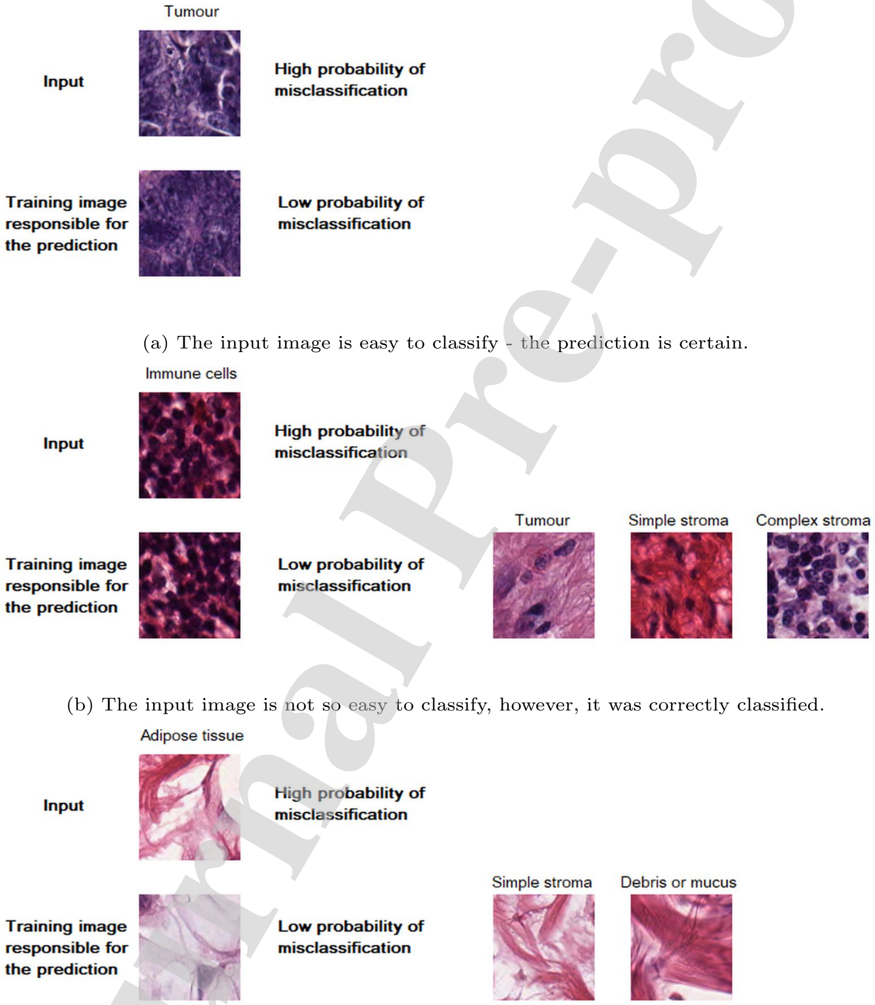
Fig. 3 shows five examples of explanations with different difficulty levels of classification of the input image. If an explanation offers tissue types with a high probability of misclassification, the input image is hard to classify. In case of only low probable misclassification offer, the input image is not easy to classify. Finally, in cases that it provides no offer, the input image is easy to classify, thus, the prediction is certain. To generate explanations, the CFCMC model trained of features extracted from *DenseNet121* CNN architecture was used.

Fig. 3a illustrates an example of a prediction explanation for the input image, which is easy to classify because it offers no other tissue types. Moreover, it can be seen that the input image is very similar to the training image responsible for the prediction (TRP). Therefore, the prediction is certain. The following semantic explanation was extracted: **The input image is for sure *Tumour epithelium*, because it could no be misclassified to any other tissue types.**

Fig. 3b shows a prediction explanation for an input image that is not so easy to classify but was correctly classified. The input image was classified as *Immune cells* tissue type. Although it offers three tissue types with a low probability of misclassification, namely *Tumour epithelium*, *Simple stroma* and *Complex stroma*, the input image is very similar to a TRP image. Therefore, the TRP image could gain trust in this prediction. The following semantic explanation was extracted: **The input image is *Immune cells* tissue type. However, there is a low possibility that in reality it could be *Tumour epithelium* or *Simple stroma* or *Complex stroma*.**

Fig. 3c shows a prediction explanation for an input image that is not easy to classify and was misclassified. The input image was predicted as *Adipose tissue type*. The explanation offers two tissue types with low probability, namely *Simple stroma* and *Debris or mucus*. It can be seen that the input image is more similar to low probable tissue types than the TRP image. This could lower trust in this particular decision. However, it offers the true tissue type of the input image, which is *Debris or mucus*. The following semantic explanation was extracted: **The input image is *Adipose* tissue type. However, there is a low possibility that in reality it could be *Simple stroma* or *Debris or mucus*.**

Fig. 3d illustrates a prediction explanation for an input image that is hard to classify, because in addition to three tissue type with a low probability for misclassification, it offers three highly probable tissue types. Therefore, the expert should investigate the input image more deeply. The following semantic explanation was extracted: **The input image is *Immune cells* tissue type. However, there is a high possibility that in reality it could be *Tumour***

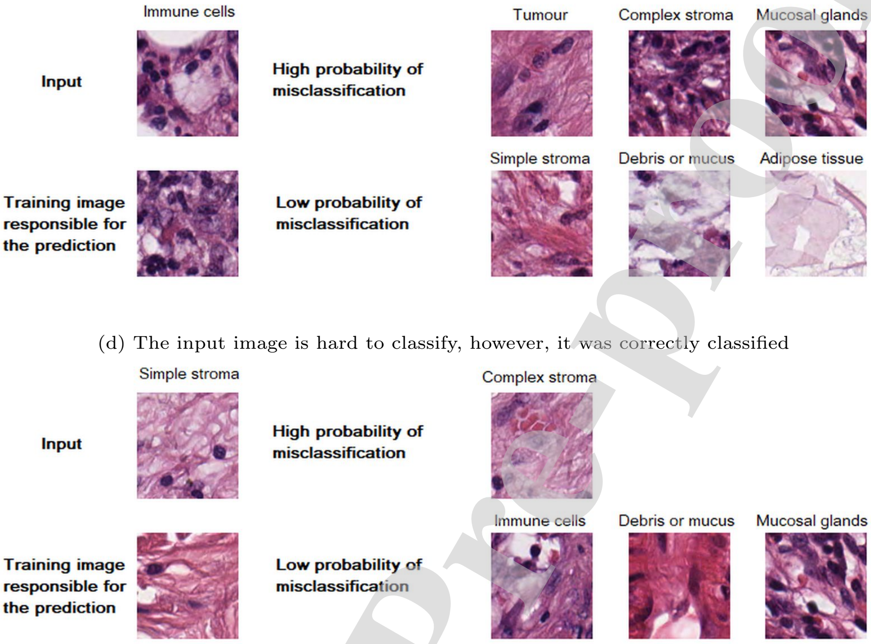

Tumour
Input
High probability of misclassification
Training image responsible for the prediction
Low probability of misclassification
Immune cells
Input
High probability of misclassification
Training image responsible for the prediction
Low probability of misclassification
Tumour
Simple stroma
Complex stroma
Adipose tissue
Input
High probability of misclassification
Training image responsible for the prediction
Low probability of misclassification
Simple stroma
Debris or mucus

(a) The input image is easy to classify - the prediction is certain.

(b) The input image is not so easy to classify, however, it was correctly classified.

(c) The input image is not so easy to classify and was wrongly classified. The true tissue type is *Debris or mucus*

Figure 3: Examples of the explanations generated for the testing dataset. The tissue types are displayed above the image.


Immune cells
Input
High probability of misclassification
Tumour
Complex stroma
Mucosal glands
Training image responsible for the prediction
Low probability of misclassification
Simple stroma
Debris or mucus
Adipose tissue
Simple stroma
Input
High probability of misclassification
Complex stroma
Training image responsible for the prediction
Low probability of misclassification
Immune cells
Debris or mucus
Mucosal glands

(d) The input image is hard to classify, however, it was correctly classified

(e) The input image is hard to classify and was wrongly classified. The true tissue type is *Complex stroma*

Figure 3: *(Continued)*

***epithelium* or *Complex stroma* or *Mucosal glands*. Moreover, there is a low possibility that it could be *Simple stroma* or *Debris or mucus* or *Adipose tissue*.**

Fig. 3e shows a prediction explanation for an input image that is hard to classify and was misclassified. It was predicted as *Simple stroma* tissue type. The explanation offers three tissue types with low and one with high probability of misclassification. It can be seen that in this case, input image is the most similar to the high probable tissue type, *Complex stroma*, which is also the true tissue type. The following semantic explanation was extracted: **The input image is *Simple stroma* tissue type. However, there is a high possibility that in reality it could be *Complex stroma*. Moreover, there is a low possibility that it could be *Immune cells* or *Debris or mucus* or *Mucosal glands*.**

5. Clinical trials results

To evaluate the influence of the explanation generated from X-CFCMC for human pathologists, we ran an acceptability test against the plain CNN. The objective of this experiment was to evaluate the acceptability of the explanation-generating X-CFCMC for human pathologists. We used within-subject experimental design, thus at the clinical trials both systems were shown to 14 pathologists (3 men and 11 women), with an an average age of 40.7 years and an average length of service of 14.9 years in which the shortest length of service was 4 years and the longest was 45 years. At the end of the session, feedback from the pathologists was collected in the form of questionnaires.

5.1. Experiment setting

Prior to the experiments, each participant was informed about the dataset. None of them were familiar with machine learning concepts. Therefore, each of the participant was informed about the automatic classification of histopathological samples by machine learning. Afterwards, both interfaces were explained to the participants, including the controls and the means of presenting the predictions. Finally, it was explained to the participants that with the exception of asking for help with controls, dialogue with the interviewers was discouraged. One different classified WSI from dataset was shown to every participant, who they were asked to examine 20 arbitrary area for both interfaces and evaluate each prediction outcome. This takes approximately 30 minutes on average. At the end of the experiment session, every participant was asked to fill out questionnaire.

5.2. Evaluation on users experiences

The users' experiences in using the stand alone CNN and X-CFCMC were evaluated using a questionnaire. The internal consistency of the questionnaire reached value of the Cronbach's alpha, $\alpha = 0.89$. Therefore, we can state that the participants sufficiently understood the objectives of the experiment. The questionnaire was divided into three parts.

The first and the second parts use a semantic differential scale, which presents respondents with a set of bipolar scales (useful/useless, reliable/unreliable). Respondents were asked to choose a number (from 1 to 6) that indicates the extent to which the adjectives relate to a characteristic evaluation of the stand alone CNN and X-CFCMC systems. While 1 represents a positive adjective (f.e. useful), 6 represents a negative adjective (f.e. useless). The adjectives were selected to cover four evaluation parameters of the systems' influences on trust and reliance to the pathologists:

1. **objectivity** - objective / subjective, useful / useless, relevant / irrelevant, serious / unserious, ethically / unethically
2. **details** - precise / imprecise, consistent / inconsistent, complicated / uncomplicated, complete / incomplete
3. **reliability** - accurate / inaccurate, faultless / faulty, straight / misleading, certain / uncertain, reliable / unreliable
4. **quality** - systematic / unsystematic, time saving / time consuming, clear / unclear, expert / inexpert, good quality / bad quality

Table 6 shows the average scores of the four evaluation parameters for both systems, which are computed as the arithmetic mean of the chosen number of each corresponding bipolar scale, while the total average is computed from the average scores of each evaluation parameter. Better values are highlighted in bold. Because the lower the score means the better the evaluation of a certain characteristic, the results reveal that the *X-CFCMC system* obtained better average scores in all parameters. The most significant differences between the scores were in the level of details (0.31) and the reliability (0.24).

The third part of the questionnaire used dichotomous scale. It consisted of closed-ended items that covered a subjective evaluation of both interfaces so that the participants express their agreement or disagreement with the statements. The statements were created to be focused on the evaluation the truthfulness and usefulness of both systems.

Analyzing the dichotomous scaled items of the third part of the questionnaires, we came to the following findings:

X-CFCMC system. The semantic explanation **influenced the trust** of 10 pathologists, while 9 of them **increased** and 1 **decreased** their trust in the prediction. *The visualization of the training image responsible for the prediction* **increases trust** of 10 of the pathologists while for the rest there was no influence. *The visualization of the other types of tissue* **influenced** only half of the pathologists, however, it increased their trust of the system.

Stand alone CNN system. The probability distribution of the prediction had no influence to 10 of the pathologists. For the rest, it increased their trust in the prediction.

Comparison of the systems. Analyzing five items devoted to a direct comparison of the usefulness of both systems, **the X-CFCMC system** achieved *a cumulative score of 50*, while **the plain CNN system** achieved *a cumulative score of 20.*

Credibility of the systems. Comparing two items about the credibility of both systems, **the plain CNN system** achieved *a cumulative score of 23*, while **the X-CFCMC system** achieved *score of 21.*

Usefulness of the whole-slide segmentation. Two items showed that, in general, all pathologists consider an automatic whole-slide segmentation of the histopathological samples useful.

5.3. Discussion

From the results above, key findings emerge. The comparison of the characteristic of both systems revealed that in the level of details, the pathologist consider the X-CFCMC system as more rigorous, more precise, more consistent, more complete. Moreover, the X-CFCMC system is considered as more accurate, reliable and confident regarding its predictions.

	Average score	
Evaluation parameters	**plain CNN**	**X-CFCMC**
Objectivity	1.75	**1.65**
Level of Details	2.16	**1.85**
Reliability	2.81	**2.57**
Quality	2.10	**1.99**
Total average	2.21	**2.01**

Table 6: A comparison of the average scores for each of the four evaluation ares for both systems, the stand alone CNN model without providing explanations and the X-CFCMC model that explains its decision

The statement evaluation indicates that the most useful means of explanations are *semantic explanation* and *a visualization of the training image responsible for the prediction. Visualization of the other types of tissue* was only appreciated by half of the pathologists. A direct comparison of both systems indicates that *the X-CFCMC system* is more acceptable than the *plain CNN system.*

6. Conclusion

In this study, we extend the explainability of the explainable Cumulative Fuzzy Class Membership Criterion (CFCMC) classifier and used it for classification of eight tissue type from histopathological cancer image samples.

First, we improved the performance of the CFCMC classifier on colorectal image data using a fine-tuned Convolutional Neural Network (CNN) as a feature extractor, which was pre-trained on different dataset.

Next, we defined the *factor of misclassification (FoM)*, which is able to estimate the possibility of the input sample being misclassified to a particular conflicting class. Moreover, we defined the *certainty threshold*, thanks to which we are able to say, whether the prediction is certain or uncertain. The proposed uncertainty measure is significantly different from many other uncertainty measures of models, e.g. neural networks model, firstly, because it is based on the classifiability of the data in the space, where the classifier makes its decision, and secondly, in the case of uncertain prediction, it is able to suggest the classes in which the input sample could be misclassified. Thus, it offers relevant classes to be further examined. The experiments clearly supported this ability.

Finally, we developed two systems for segmentation of the whole-slide images of histopathological cancer tissue. The first system used stand alone CNN and the second used an X-CFCMC classifier, which provides three means of explanations: a semantical explanation about the prediction and possible misclassification, a visualization of the training image responsible for the prediction and a visualization of the other types of tissue.

At the clinical trials with 14 pathologists, we measured the acceptability and the trust of the pathologists for the proposed system. The results indicate that the X-CFCMC system is more useful and more reliable than the plain CNN.

In conclusion, this paper discussed about the usability and the reliability of a explainable classifier in real world medical settings through clinical trials. We believe that our proposed system can contribute to the use of AI, especially by improving the usability and acceptability of AI systems in medical domains, where speed in making decisions, reliability and accountability are crucial. We are aware that the scale of our preliminary experiment was limited. The expansion of clinical trials to include more pathologists from various fields are of our immediate future interest. Moreover, in the medical settings we will be confronted with imbalanced, heterogeneous and inaccurate data sets. Therefore, our next challenge in our research is also to examine how our classifier will perform on imperfect data.

Acknowledgment

This research is supported by AI4EU project from the European Union's Horizon 2020 research and innovation programme under grant agreement 825619 2019-2021, Maria Currie RISE LIFEBOTS Exchange Grant Agreement ID : 824047 2019-2021 and EU FlagEra Joint project Robocom++, 2017-2021.

References

[1] J. Xie, R. Liu, I. Luttrell, C. Zhang, et al., Deep learning based analysis of histopathological images of breast cancer, Frontiers in genetics 10 (2019) 80.

[2] J. N. Kather, C.-A. Weis, F. Bianconi, S. M. Melchers, L. R. Schad, T. Gaiser, A. Marx, F. G. Zöllner, Multi-class texture analysis in colorectal cancer histology, Scientific reports 6 (2016) 27988.

[3] B. E. Bejnordi, G. Zuidhof, M. Balkenhol, M. Hermsen, P. Bult, B. van Ginneken, N. Karssemeijer, G. Litjens, J. van der Laak, Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images, Journal of Medical Imaging 4 (4) (2017) 044504.

[4] J. de Matos, A. d. S. Britto Jr, L. E. Oliveira, A. L. Koerich, Histopathologic image processing: A review, arXiv preprint arXiv:1904.07900.

[5] D. Komura, S. Ishikawa, Machine learning methods for histopathological image analysis, Computational and structural biotechnology journal 16 (2018) 34–42.

[6] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, A. Campilho, Classification of breast cancer histology images using convolutional neural networks, PloS one 12 (6) (2017) e0177544.

[7] M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, A. Binder, Resolving challenges in deep learning-based analyses of histopathological images using explanation methods, arXiv preprint arXiv:1908.06943.

[8] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv preprint arXiv:1712.09923.

[9] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Mueller, Causability and explainability of ai in medicine, Data Mining and Knowledge Discovery doi 10.

[10] A. Holzinger, M. Plass, M. Kickmeier-Rust, K. Holzinger, G. C. Crişan, C.-M. Pintea, V. Palade, Interactive machine learning: experimental evidence for the human in the algorithmic loop, Applied Intelligence 49 (7) (2019) 2401–2414.

[11] G. R. Vásquez-Morales, S. M. Martínez-Monterrubio, P. Moreno-Ger, J. A. Recio-García, Explainable prediction of chronic renal disease in the colombian population using neural networks and case-based reasoning, IEEE Access 7 (2019) 152900–152910.

[12] J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein, Explainable prediction of medical codes from clinical text, arXiv preprint arXiv:1802.05695.

[13] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nature biomedical engineering 2 (10) (2018) 749.

[14] A. Malhi, T. Kampik, H. Pannu, M. Madhikermi, K. Främling, Explaining machine learning-based classifications of in-vivo gastral images, in: 2019 Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2019, pp. 1–7.

[15] P. Hartono, A transparent cancer classifier, Health Informatics Journal 26 (1) (2020) 190–204, pMID: 30596318. doi:10.1177/1460458218817800.

[16] Ł. Raczkowski, M. Możejko, J. Zambonelli, E. Szczurek, Ara: accurate, reliable and active histopathological image classification framework with bayesian deep learning, bioRxiv (2019) 658138.

[17] P. Sabol, P. Sinčák, K. Ogawa, P. Hartono, Explainable classifier supporting decision-making for breast cancer diagnosis from histopathological images, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[18] P. Sabol, P. Sinčák, J. Buša, P. Hartono, Cumulative fuzzy class membership criterion decision-based classifier, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017, pp. 334–339. doi:10.1109/SMC.2017.8122625.

[19] P. Sabol, P. Sinčák, J. Magyar, P. Hartono, Semantically explainable fuzzy classifier, International Journal of Pattern Recognition and Artificial Intelligence 33 (12) (2019) 2051006. arXiv:https://doi.org/10.1142/S0218001420510064, doi:10.1142/S0218001420510064.
URL https://doi.org/10.1142/S0218001420510064

[20] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1885–1894.

[21] J. Zhou, H. Hu, Z. Li, K. Yu, F. Chen, Physiological indicators for user trust in machine learning with influence enhanced fact-checking, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, 2019, pp. 94–113.

[22] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber, et al., Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study, PLoS medicine 16 (1) (2019) e1002730.

[23] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

[25] S. Mohseni, N. Zarei, E. D. Ragan, A survey of evaluation methods and measures for interpretable machine learning, arXiv preprint arXiv:1811.11839.

[26] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63 (2) (2001) 411–423.

[27] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.

[28] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE transactions on pattern analysis and machine intelligence (2) (1979) 224–227.

[29] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, MIT press, 1992.

[30] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[31] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2818–2826.

[32] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[33] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.

[34] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[35] C. Szegedy, S. Ioffe, V. Vanhoucke, A. A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Thirty-first AAAI conference on artificial intelligence, 2017.

[36] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, arXiv preprint arXiv:1905.11946.

[37] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980.

***Credit Author Statement**

Patrik Sabol – Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – Original Draft, Writing – Review & Editing, Data curation
Peter Sinčák – Conceptualization, Supervision, Funding acquisition, Project administration
Pitoyo Hartono – Conceptualization, Methodology, Supervision, Writing – Original Draft
Pavel Kočan – Conceptualization, Investigation, Resources
Zuzana Benetinová – Investigation, Resources, Data curation
Alžbeta Blichárová – Investigation, Resources
Ľudmila Verbóová – Investigation, Resources, Writing – Review & Editing
Erika Štammová – Investigation, Resources, Validation
Antónia Fabianová-Sabolová – Formal analysis, Methodology, Visualization, Data curation
Anna Jašková - Formal analysis, Methodology, Visualization

Declaration of interests

☒ 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.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: