Communication Papers of the of the 17th Conference on Computer Science and Intelligence Systems pp. 23–30
DOI: 10.15439/2022F128
ISSN 2300-5963 ACSIS, Vol. 32

Detecting Cancerous Regions in DCE MRI using Functional Data, XGboost and Neural Networks

Povilas Treigys, Jolita Bernatavičienė
Vilnius University
Institute of Data Science and Digital Technologies
Akademijos st. 4, Vilnius, Lithuania
Email: {povilas.treigys, jolita.bernataviciene}@mif.vu.lt

Aleksas Vaitulevičius
Vilnius University
Institute of Data Science and Digital Technologies
Akademijos st. 4, Vilnius, Lithuania
Email: aleksas.vaitulevicius@mif.stud.vu.lt

Jurgita Markevičiūtė
Vilnius University
Institute of Applied Mathematics
Naugarduko st. 24, Vilnius, Lithuania
Email: jurgita.markeviciute@mif.vu.lt

Ieva Naruševičiūtė, Mantas Trakymas
Vilnius University
National Cancer Institute
Santariškių st. 1, Vilnius, Lithuania
Email: {ieva.naruseviciute, mantas.trakymas}@nvi.lt

***Abstract*—Cancerous region detection in the prostate is performed using different imaging sequences by multiparametric magnetic resonance imaging. One of those modalities is dynamic contrast enhancement. The authors of this paper are testing possible modifications of workflow which use this modality for more accurate cancerous region detection in the prostate. The introduced changes are timestamp mapping in the segmentation step, proportionate Simple Linear Iterative Clustering region number to prostate region size in each slice, a new definition of labels and new extracted features. Furthermore, experiments are performed for segmentation in a single timestamp only. The experiments test the effect of modification on curve classification by using XGBoost classification and flat neural network approaches. Lastly, the authors perform hyper-parameter tuning of both approaches and evaluates obtained results statistically.**

I. Introduction

One of the most lethal cancer globally is prostate cancer. According to the research by Bray et al. in paper [6], it has the second highest incidence rate among males after lung cancer. Successful prostate cancer treatment requires early diagnosis. Preliminary identification of cancer is related to a higher concentration of a protein produced by the prostate called Prostate-Specific Antigen (PSA), as it is described in paper [17] by Hayes and Barry. However, PSA testing has a high level of false-positive and false-negative cases. Therefore, in addition to PSA testing, many biopsies are taken, which is a highly invasive testing method. As an alternative to this method, PI-RADS is introduced in paper [2]. It is a structured reporting scheme for multiparametric (mp) prostate Magnetic Resonance Imaging (MRI). Literature evidence of the same paper and the expert opinion consensus indicate better interpretation and performance of prostate cancer evaluation when using PI-RADS rather than PSA testing. Examples of MRI modalities used for cancer evaluation are T2 weighted images (T2W), Diffusion Weighted Images (DWI), Apparent Diffusion Coefficients (ADC), and Dynamic Contrast-Enhanced (DCE) images.

This work was not supported by any organization

Vaitulevicius et al. introduce in paper [27] the workflow for detecting cancerous regions in the prostate by using DCE sequences together with preliminary research. This modality is acquired by capturing a sequence of MRI scans during intravenous injection of a contrast agent. Typically gadolinium is used, and the scans are performed every few seconds for several minutes. As described by Low et al. in paper [23], during this period, tumours attract a higher amount of contrast medium due to their typically higher vascular permeability and density caused by angiogenesis. This data can detect, characterize, and monitor tumours using Functional Data Analysis (FDA) and machine learning methods. The approach tested in the experiments described by Vaitulevičius et al. in paper [27] performs the following steps. Firstly, these DCE sequences' cross-sectional images are segmented using image segmentation algorithms such as the Simple Linear Iterative Clustering (SLIC) algorithm. Secondly, the regions are aggregated to values representative of regions by calculating such metrics as mean or median. Thirdly, for each region x function $f_x : T \to I$ are created by fitting aggregated values of region x. In the definition of these functions T is the set of timestamps while I is the set of aggregated values at the timestamps T. Finally, the preliminary research is performed on the K-nearest neighbours algorithm (KNN) using functional data and the support vector machine (SVM) algorithm using extracted features from functional data.

In this paper, several adjustments to the workflow are tested. Firstly, instead of a fixed number of SLIC regions, using a proportionate number of SLIC regions in each slice is proposed and compared to the workflow described in paper [27]. Secondly, the experiments are done to determine if using landmark registration on data improves or worsens the classification accuracy metrics. Moreover, the experiments are repeated on thirteen patients separately instead of one.

Furthermore, this paper focuses entirely on segmenting each timestamp individually and does not introduce any experiments on the temporal variance matrix calculated between all timestamps. This paper chooses only a single timestamp for each patient for segmentation. Finally, two new algorithms, flat neural network and XGBoost classification introduced in paper [11], are tested using features extracted from functional data as these algorithms are more advanced and highly tunable. Therefore, the hyperparameters of those algorithms are tuned. The data used for experiments is the same as in the experiments introduced in paper [27]. The data for the investigation, under the terms of the bioethical agreement, was provided by Lithuanian National Cancer Institute.

A lot of prostate cancer research with MRI data is already done. Many examples of machine learning applications on MRI modalities are given in paper [13]. Papers [19] and [1] conducted research tests the possibility of using T2W sequences for cancer localization. Another examples are usage of DWI sequences to solve prostate cancer segmentation and severity evaluation problems presented in papers [26], [18], [29] and [4]. Moreover, in paper [14] research was conducted which tested the capability of various adaptations of U-net to detect and grade cancerous tissue by using T2-weighted and DWI modalities. Lastly, technological improvements such as the ones presented in paper [10] in DCE MRI modality data acquisition creates a demand to research DCE MRI sequences. One of the research on DCE MRI sequences was performed in paper [20]. However, the paper [20] focuses purely on machine learning and does not use FDA approach.

To summarize, authors of this paper introduced and tested adjustments to the workflow described in article [27] by Vaitulevičius et al. The changes are segmentation timestamp mapping, using a proportionate number of SLIC regions to prostate size in each slice, label definition, extracted features and new classification models.

II. Experiment setup

As mentioned in the introduction, several adjustments to the workflow described in paper [27] was introduced. The experiments are performed according to a new workflow. Data used in the experiments is the same as in the experiment of previously published author's investigation. However, investigations of this paper, differently from the experiment described in article [27], are performed on 13 patients instead of one by performing model training and validation for each patient separately. The example of a single patient's single slice in 3 different timestamps is visualized in Fig. 1. The left shows the DCE image taken at the timestamp at which the contrast agent has not yet reached the prostate region. The middle shows the DCE image is taken at the timestamp the contrast agent is flowing through arteries, and the right DCE MRI image is taken at the timestamp when the contrast agent is accumulating in the prostate. Meanwhile, in Fig. 2 cancerous and prostate masks are displayed. Lastly, each patient's data is split into a training set of 70% of the patient's dataset and a validation set of 30% of the patient's dataset. The ratio of classes in training and validation data sets is equal.

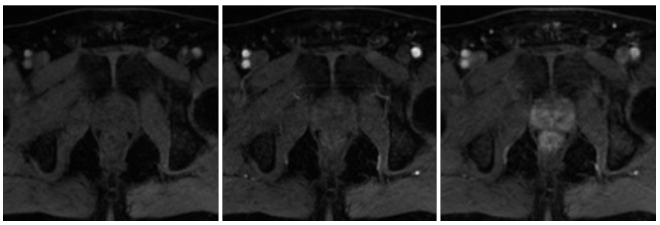

Fig. 1. Example of the DCE images of a single slice at 3 different timestamps.

Flat neural network and XGBoost algorithms are tested in the experiments presented in this paper. Authors test only the simple flat neural network, consisting of only dense layers and compare it to another machine learning algorithm - XGBoost. Flat neural network is chosen as a baseline as deep neural networks are highly scalable group of machine learning methods and flat neural network is the most simple one. Meanwhile, XGBoost is a relatively new algorithm which has already proven to be effective in solving various tasks in the medical field. For example, the experiment provided in paper [9] indicates that XGBoost achieves the most accurate results when predicting the outcome of hypertension. The other example is the experiment provided in article [16]. Authors show that XGBoost excels in the pathway analysis, which is used to determine the role of the tested protein task. One more example is the experiment provided in paper [21], which indicates that XGBoost accurately predicts mortality from acute kidney injury.

A. Chose of timestamps for each patient

Each patient's data is acquired at different timestamps and do not correspond to each other. For example, if patient A has 31 timestamps while patient B has 20 timestamps, then the patient's A timestamp five will not conform to patient's B timestamp five. Meanwhile, chosen timestamp for segmentation for each patient has to be conforming in order to not introduce the effect of the timestamp choice. To overcome this problem, each patient's timestamps are mapped to 25 timestamps. An illustrative scheme of mapping is displayed in Fig. 3 where columns Patient A timestamps and Patient B timestamps correspond to original timestamps. Firstly, the timestamps of the patients are normalized to the interval

Fig. 2. The left image is a cancerous region mask and the right image is the prostate mask of a single slice.

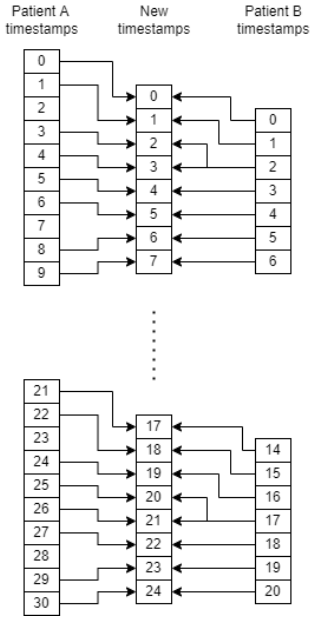

Fig. 3. Example of the timestamp mapping for Patient A data having 31 timestamps and Patient B data having 20 timestamps.

[0, 24]. Secondly, normalized timestamps are rounded to the whole numbers. The resulting new timestamps are displayed in the column New timestamps. However, after this step, patients with more than 25 timestamps have multiple DCE images at a single timestamp, while patients with less than 25 timestamps have none at specific timestamps. Therefore, if more than one patient's timestamp is mapped to a new timestamp, then the latest timestamp of those timestamps is selected for a new timestamp. Suppose none of the patient's timestamps gets mapped to a new timestamp. In that case, patient's normalized rounded timestamp, which has a minor absolute difference with a new timestamp, is mapped to a new timestamp. If more than one of such timestamps exists, then the earliest timestamp is mapped to a new timestamp. As mentioned in the introduction, a single timestamp is used for experiments when performing segmentation. The chosen timestamp was the 5th.

B. Segmentation

In the original segmentation described in article [27], each patient's slice is segmented into a fixed number of zones, 50. However, it seems it is biased towards SLIC regions of slices in which prostate regions are smaller as they have a smaller area, but the number of regions produced from them is the same.

This paper investigates the ability to select the number of SLIC regions proportionate to the size of the slice. Firstly, the slice with the largest prostate region is chosen. For that slice, 50 of SLIC regions are obtained. Secondly, number of SLIC regions in other slices are calculated by using formula $n_i = n_{max} \times s_i / max(S)$ where n_{max} is number of SLIC regions to which the slice with largest prostate region is segmented, s_i - prostate size in the slice, S - list of prostate region sizes across slices. The results are rounded to whole numbers. These results are the amount of SLIC regions for each slice. The difference between the original and modified approaches is illustrated in Fig. 4.

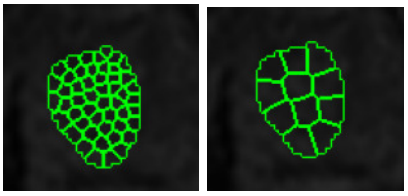

Fig. 4. Example of the slice segmentation. On the left the result of modified segmentation approach, on the right - the original segmentation is displayed.

C. SLIC region labels

In paper [27], positive class was assigned to the SLIC regions, which had an overlap of $\geq 50\%$ with cancer mask and with malignant biopsy mask. However, this definition resulted in a very small amount of samples of regions with positive class labels resulting in an imbalanced class problem. Therefore in this paper, a new description of positive class label is applied, which results in less accurate labels, but it increased the data sample amount for positive class labels.

In the research described in this paper, positive class is assigned to SLIC regions with the overlap of $\geq 50\%$ with cancer mask, which has overlap with at least one malignant biopsy mask and no overlap with benign biopsy masks. SLIC regions to which negative label 0 is assigned remain the same as in paper [27]. The rest of the SLIC regions are not used in training or validation. These SLIC regions include regions that either have:

- Overlap of $< 50\%$ with cancer masks, which overlap with at least one malignant biopsy mask and has no overlap with benign biopsy masks.
- Overlap with cancer mask, which has no overlap with malignant biopsy mask.

The example of SLIC region labels defined in this paper is displayed in Fig. 5.

D. Features from functional data

A different set of features are chosen for the training and validation in the experiments described in this paper than in the experiments described by Vaitulevičius et al. in paper [27]. Firstly, the number of uniformly spaced discrete values is increased from 10 to 100. Such adjustment results in the longer training process and increased accuracy of functional data representation. Secondly, the maximal value's timestamp normalized to the interval [0, 1] is not used for the experiments as it is the same for the registered functional data. After registering the data, this feature becomes equal for all regions. Therefore, it does not carry any valuable information for experiments conducted with functional data registration. Thus, extracted features in the investigations of this paper are:

- uniformly spaced discrete values - $f(x_i)$ where f is a single function from functional data and $x_i \in [0, 1]$, $x_i - x_{i-1} = x_j - x_{j-1}$ with $i \in [0, 100]$ and $j \in [0, 100]$.
- max value of the function - $\sup_{x\in[0,1]} f(x)$.
- modified band depth - functional depth described in paper [22].
- integral depth - functional depth described in paper [15].

Lastly, two types of experiments are conducted. First experiment uses only discrete FDA values. Second - discrete values obtained from FDA and the features such as: modified band depth, integral depth and max value of the function. If the second experiment will not bear significant results, then that would indicate that no additional information is needed than the discrete values of the functional data itself

E. Software used for experiments

All experiments are performed by using Python 3.8 programming language with latest packages:

- scikit-fda - smoothing time series into functional data and functional data transformations.
- xgboost - XGBoost classification model training. The library is introduced in paper [11]
- hyperopt - XGBoost classification model's hyperparameter tuning. The library is introduced in paper [8].
- keras - flat neural network model's training with tensorflow framework. The library is published in [12].
- keras_tuner - flat neural network model's hyperparameter tuning. The library is published in [24].
- scikit-learn - data splitting and metric calculation. The library is introduced in paper [25].
- scipy - statistical tests.

III. Hyperparameter tuning

Hyperparameter tuning of both algorithms is performed using each patient's data separately. Tuning allows the investigation of possible hyperparameters for a more general model. Persistent values of hyperparameters indicate that those values can be used in a more general model. However, non-persistent values suggest that the model is too simple and hardly can be generalized.The dataset is split into stratified training and validation sets for hyperparameter tuning. The validation set contains 30% of the dataset while the training set - 70%. Each feature of the training set is scaled by normalizing it to the interval [0, 1], while validation set features are scaled by dividing each by the maximum value of the training set.

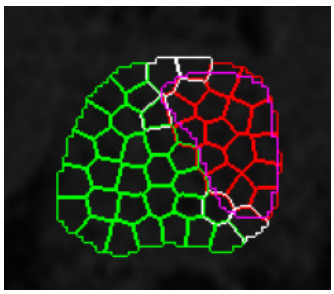

Fig. 5. Example of SLIC region labels. The regions with a red contour are regions representing the positive class label. The regions with a green outline are regions representing the negative class label. The labels with white contours are not used in the training or validation. The purple contour is the cancerous region mask which has overlap with malignant biopsy mask and no overlap with benign biopsy masks

Moreover, due to the high class imbalance, additional class weights are assigned. For positive class weight is calculated by using formula $w_p = 1 - (n_p/n_a)$ and for negative class - $w_n = 1 - (n_n/n_a)$ where n_p is the number of SLIC regions with a positive class in the training set, n_n is the number of SLIC regions with a negative class in the training set, n_a - size of the training set.

Lastly, early stopping is used for XGBoost classification and flat neural network algorithms to avoid overfitting. For XGBoost, if after iterating through the training set ten times F1 score of the validation set does not improve, then the training is stopped. For a flat neural network, if after ten epochs F1 score of the validation set does not improve, then the training is stopped, and the model with the highest F1 score of the validation set is chosen.

A. XGBoost algorithm hyperparameter tuning

Hyperparameters of the XGBoost classification algorithm are tuned by using Tree of Parzen Estimators (TPE) described by Bergstra et al. in paper [7]. The training for each patient is repeated at least 400 times in hyper-parameter tuning. Hyperparameters are tuned by maximizing the F1 score on the validation data set. The space of tuned hyperparameters is taken from Kagle competition [3]. Tuned hyperparameters are:

- Subsample ratio of columns when constructing each tree. The search space of this hyperparameter is in interval [0.4, 0.8].
- Minimum loss reduction required to make a further partition on a leaf node of the tree. The search space of this hyperparameter is in interval [0, 1].
- Maximal depth of the XGBoost ensemble tree. The search space of this hyperparameter is in an interval of natural numbers [3, 18].
- Minimum sum of instance weight (Hessian) needed in a child. The search space of this hyperparameter is in an interval of natural numbers [0, 10].
- L1 regularization term on weights. The search space of this hyperparameter is in interval [0, 1].

Number of gradient boosted trees. This hyperparameter is a constant and set to 180 and each training is performed by minimizing the logistic regression loss function.

B. Flat neural network algorithm hyperparameter tuning

Hyperparameters of the flat neural network algorithm are tuned by using the Bayesian tuning algorithm described by Barsce et al. in paper [5]. The training for each patient is repeated at least 25 times. Each training is performed in 100 epochs or less (if the early stop is triggered). Due to the non-deterministic behaviour of the flat neural network training

algorithm, each hyperparameter tuning procedure has been repeated ten times. Hyperparameters are tuned by maximizing the mean of the F1 score on the validation data set. The architecture of the tuned flat neural network is:

- Input layer
- First hidden layer is a dense layer with a ReLU activation function and the number of neurons equal to a number of input dimensions.
- Search the number of hidden dense layers (depth of neural network). The search space of this hyperparameter is in an interval of natural numbers [1, 10]. Each layer's activation function is ReLU. A number of neurons are also tuned and is set to repetitive numbers of 32 in the interval of natural numbers [32, 1024].
- Last layer of the neural network contains only a single neuron. The activation function of this layer is tuned by testing activation functions - sigmoid, softmax and ReLU.

IV. Results

In the experiments provided in this paper, hyperparameter tuning and training are done on each patient's data separately. The collected accuracy metrics of the resulting models are calculated on the validation set. Accuracy metrics are precision, recall, F1, balanced accuracy and specificity. Experiments with the flat neural network are repeated ten times with each patient's dataset. Collected accuracy metrics of flat neural network's model are aggregated on a patient and configuration basis by calculating the mean and the standard deviation. Those metrics can be used to choose a more stable model as mean is heavily affected by outliers and standard deviation indicates how much do the outliers affect the mean. Lastly, the median value of each accuracy metric is calculated on configuration basis as median have lesser effect of outliers and represent the sample distribution better than mean. Those configurations are registered, unregistered functional data, only discrete and not only discrete extracted features from functional data, proportionate and fixed number of SLIC regions in the slice and lastly selected model - XGBoost or flat neural network.

Tables I and III contain medians of validation dataset classification accuracy metrics of flat neural network and XGBoost model respectively. Tables II and III contain hyperparameters acquired by hyperparameter tuning on dataset of single patient for flat neural network and XGBoost model respectively. The shown hyperparameter tuning result is acquired from the patient's validation dataset, which is classified with the highest balanced accuracy.

In the tables I, II, and III is registered column denotes if functional data in the experiments is registered (registered) or not (unregistered). Column extracted features denotes if only discrete values of functional data are used (only discrete) or discrete values, maximal values, integral depth, or modified band depth (not only discrete). Column number of SLIC regions denotes if, in the experiments, segmentations in each slice are performed with a fixed number of SLIC regions (fixed SLIC) or not (proportionate SLIC).

Table III indicates that the highest median of balanced accuracy - 0.855 is achieved by using non-registered functional data, a proportionate number of SLIC regions to prostate size and extracted features: discrete values, modified band depth, integrated depth and maximal values. For this configuration, tuned hyperparameters on the patient's dataset with the best balanced accuracy were achieved is:

- Subsample ratio of columns when constructing each tree - 0.412 (column colsample bytree)
- Minimum loss reduction required to make a further partition on a leaf node of the tree - 0.005 (column gamma)
- Maximum tree depth for base learners - 6 (column max depth)
- Minimum sum of instance weight(hessian) needed in a child - 0 (column min child weight)
- L1 regularization term on weights - 0.53 (column reg alpha)

Table I indicates that the highest median of mean balanced accuracy - 0.831 is achieved using non-registered functional data, a fixed number of SLIC regions and extracted features: discrete values, modified band depth, integrated depth and

TABLE I
MEDIANS OF FLAT NEURAL NETWORK MODEL'S CLASSIFICATION ACCURACY METRICS ON VALIDATION DATASET

is registered	extracted features	number of SLIC regions	precision mean	precision std	recall mean	recall std	f1 mean	f1 std	**balanced accuracy mean**	balanced accuracy std	specificity mean	specificity std
unregistered	not only discrete	proportionate SLIC	0.369	0.061	0.650	0.071	0.442	0.046	0.772	0.031	0.923	0.036
		fixed SLIC	0.279	0.052	0.733	0.084	0.391	0.040	**0.831**	0.037	0.943	0.028
	only discrete	proportionate SLIC	0.353	0.074	0.633	0.091	0.413	0.059	0.784	0.044	0.936	0.037
		fixed SLIC	0.332	0.041	0.738	0.114	0.395	0.043	0.812	0.056	0.942	0.016
registered	not only discrete	proportionate SLIC	0.332	0.087	0.642	0.145	0.393	0.077	0.775	0.060	0.928	0.030
		fixed SLIC	0.181	0.061	0.660	0.141	0.270	0.044	0.772	0.067	0.935	0.036
	only discrete	proportionate SLIC	0.282	0.078	0.686	0.113	0.364	0.077	0.782	0.052	0.911	0.036
		fixed SLIC	0.186	0.047	0.667	0.076	0.268	0.038	0.801	0.021	0.933	0.017

TABLE II
RESULTS OF FLAT NEURAL NETWORK MODEL'S TUNING (DISPLAYED HYPERPARAMETERS ARE ONLY OF ONE PATIENT FOR WHICH BALANCED ACCURACY ON THE VALIDATION SET IS THE HIGHEST)

is registered	extracted features	number of SLIC regions	activation function	0	1	2	3	4	5	6	7	8	9	10
unregistered	not only discrete	proportionate SLIC	sigmoid	864	416									
		fixed SLIC	sigmoid	1024										
	only discrete	proportionate SLIC	sigmoid	1024										
		fixed SLIC	sigmoid	1024										
registered	not only discrete	proportionate SLIC	sigmoid	1024	384	320								
		fixed SLIC	sigmoid	1024										
	only discrete	proportionate SLIC	sigmoid	1024	32	32	32	32	32	32	32	480	32	1024
		fixed SLIC	sigmoid	672	32	32	32	32	32	32	32	32		

maximal values. Table II indicates that for this configuration, tuned hyperparameters on the patient's dataset with the best balanced accuracy were achieved using the activation function of the last layer - Sigmoid (column activation function) and one layer with 1024 neurons.

The table IV accuracy metrics are aggregated to a single configuration basis by calculating medians. Those configurations are:

- if functional data used in the experiment is registered (registered) or not (unregistered).
- if only discrete values extracted from functional data are used (only discrete), or the addition of discrete values such as modified band depth, integral depth and max values (not only discrete) increases accuracy.
- if in the experiments, segmentations in each slice are performed with a fixed number of SLIC regions (fixed SLIC) or proportionate number of SLIC regions to prostate size (proportional SLIC).
- if the experiments are performed with XGBoost classification (XGBoost) or flat neural network (flat neural network).

Table IV indicates that using unregistered function data (balanced accuracy - 0.804) gives more accurate results than using registered functional data (balanced accuracy - 0.773). Furthermore, it also indicates that using only extracted discrete values from functional data (balanced accuracy - 0.789) gives slightly more accurate results than using maximal values, integrated depth and modified band depth (balanced accuracy - 0.78). However, that contradicts the one made from tables I and III. This contradiction indicates that other configurations have effect on how well do extracted features perform or that the effect of this configuration is very small. Moreover, table IV indicates that using a fixed number of SLIC regions (balanced accuracy - 0.792) gives more accurate results than using a proportionate number of SLIC regions to prostate size (balanced accuracy - 0.783). However, tables I and III indicate that this configuration is dependent on which machine learning algorithm is used. Lastly, the table IV indicates that XGBoost classification algorithm (balanced accuracy - 0.792) is better than Flat neural network (balanced accuracy - 0.785).

Finally, statistical tests are performed on acquired balanced accuracy values. A single sample is formed for each configuration by taking all balanced accuracies that use that exact configuration. Further paired statistical tests are performed on sample pairs of configurations which are compared in the experiments of this paper. As samples are relatively small, the chosen tests are Wilcoxon tests introduced in paper [28]. The resulting p-values acquired by performing statistical tests on samples of balanced accuracies are as follows:

- Using a flat neural network and using XGBoost model -

TABLE III
MEDIANS OF XGBOOST MODEL'S CLASSIFICATION ACCURACY METRICS ON VALIDATION DATASET AND RESULTS OF XGBOOST MODEL'S TUNING (DISPLAYED HYPERPARAMETERS ARE ONLY OF ONE PATIENT FOR WHICH BALANCED ACCURACY ON THE VALIDATION SET IS THE HIGHEST

is registered	extracted features	number of SLIC regions	precision	recall	f1	**balanced accuracy**	specificity	colsample bytree	gamma	max depth	min child weight	reg alpha
unregistered	not only discrete	proportionate SLIC	0.400	0.750	0.471	**0.855**	0.959	0.412	0.005	6	0	0.530
		fixed SLIC	0.500	0.600	0.540	0.794	0.985	0.458	0.002	15	0	0.636
	only discrete	proportionate SLIC	0.455	0.667	0.462	0.810	0.964	0.420	0.000	8	0	0.596
		fixed SLIC	0.500	0.640	0.514	0.793	0.971	0.457	0.247	13	0	0.008
registered	not only discrete	proportionate SLIC	0.571	0.650	0.533	0.780	0.971	0.727	0.019	16	0	0.825
		fixed SLIC	0.450	0.560	0.483	0.748	0.976	0.665	0.080	13	1	0.446
	only discrete	proportionate SLIC	0.500	0.625	0.429	0.752	0.977	0.586	0.020	6	0	0.046
		fixed SLIC	0.500	0.667	0.465	0.750	0.980	0.698	0.007	10	0	0.505

TABLE IV
MEDIANS OF ACCURACY METRICS OBTAINED ON THE VALIDATION DATASET. THIS AGGREGATION IS PERFORMED ON CONFIGURATION BASIS (THE CONFIGURATION IS DENOTED WITH COLUMN EXPERIMENTS DONE WITH)

configuration	precision	recall	f1	**balanced accuracy**	specificity
unregistered	0.382	0.675	0.446	0.804	0.951
registered	0.361	0.657	0.431	0.773	0.948
not only discrete	0.389	0.667	0.457	0.780	0.948
only discrete	0.356	0.667	0.426	0.789	0.951
proportionate SLIC	0.394	0.667	0.450	0.783	0.948
fixed SLIC	0.355	0.667	0.431	0.792	0.953
XGBoost	0.500	0.652	0.502	0.792	0.973
flat NN	0.303	0.675	0.366	0.785	0.934

0.289765.
- Using registered functional data and unregistered functional data - 0.000007.
- Using additional extracted features and using discrete values only - 0.203341.
- Using a fixed number of SLIC regions and a proportionate number of SLIC regions - 0.502053.

To summarize the obtained results, the following comparisons are performed:

- XGboost model's results are compared to flat neural network's results:
  - Median of all validation dataset classification balanced accuracy metrics by using XGBoost classification algorithm is 0.792, while using a flat neural network - 0.785.
  - The highest median of validation dataset balanced accuracy calculated for each configuration separately with XGBoost classification algorithm is 0.855 while with the flat neural network - 0.831.
  - P-value of statistical test performed between these balanced accuracies is 0.289765.
- Using unregistered functional data is compared with registered functional data usage:
  - Median of all validation dataset classification balanced accuracy metrics using non-registered functional data is 0.804, while registered functional data is 0.773.
  - The highest median of validation dataset balanced accuracy calculated for each configuration separately with non-registered functional data is 0.855 while with registered functional data - 0.801.
  - P-value of statistical test performed between these balanced accuracies is 0.000007.
- Using additional extracted features from functional data is compared with using a dataset which represents functional data itself only:
  - Median of all validation dataset classification balanced accuracy metrics by using additional extracted features from functional data is 0.780, while dataset which represents functional data itself only - 0.789.
  - The highest median of validation dataset balanced accuracy calculated for each configuration separately with additional extracted features from functional data is 0.855 while with a dataset representing functional data itself only - 0.812.
  - P-value of statistical test performed between these balanced accuracies is 0.203341.
- Using a proportionate number of SLIC regions to prostate size is compared with using a fixed number of SLIC regions:
  - The highest median of validation dataset balanced accuracy calculated for each configuration separately by using a proportionate number of SLIC zones to prostate size is 0.855 while a fixed number of SLIC zones - 0.794.
  - Using flat neural network algorithm as the highest median of validation dataset balanced accuracy calculated for each configuration separately by using a proportionate number of SLIC zones to prostate size is 0.784 while a fixed number of SLIC zones - 0.831.
  - P-value of statistical test performed between these balanced accuracies is 0.502053.

V. CONCLUSION

The results obtained by this research that:

- XGBoost classification algorithm gives slightly more accurate results (in configuration investigation as well) than the flat neural network. However, the obtained better performance is insignificant as the p-value of the statistical test performed between obtained results is 0.29.
- Unregistered functional data gives significantly more accurate results than registered (in configuration investigation as well). The median of classification balanced accuracy metrics using non-registered functional data is 0.804, while registered functional data - 0.773. The result difference is statistically significant as the p-value of the statistical test performed between these balanced accuracies is less than 0.05.
- There is no significant difference in classification results between using additional extracted features from functional data and using a dataset which represents functional data only (in configuration investigation as well). The median of classification balanced accuracy metrics using additional extracted features from functional data is 0.78, while dataset which represents functional data only - 0.789. However, this comparison is insignificant as the

p-value of the statistical test performed between these balanced accuracies is 0.203.

- Application of proportionate number of SLIC zones to prostate size gives more accurate results than a fixed number of SLIC zones when using XGBoost classification algorithm. The highest median of balanced accuracy calculated for each configuration by using a proportionate number of SLIC zones to prostate size is 0.855 while a fixed number of SLIC zones - 0.794. In opposite, the flat neural network algorithm performs better with fixed number of SLIC zones - 0.831, that proportionate - 0.784 (in terms of balanced accuracy and configuration investigation). This comparison is insignificant as the p-value of the statistical test performed between these balanced accuracies is 0.502.

Results obtained indicate further research directions:

- The experiments should be repeated on higher data variability from more patients. The data variability could be used to explain the proportionate number SLIC zones performance with flat neural networks.
- The search of ensemble classifier that merge the proposed scheme of processing DCE modality with processing other prostate MRI modalities could improve the results.

ACKNOWLEDGMENT

The authors are thankful for the high performance computing resources provided by the Information Technology Research Center of Vilnius University.

REFERENCES

[1] R. Alkadi, F. Taher, A. El-Baz, and N. Werghi, "A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images," *Journal of digital imaging*. 32(5), (2019):793-807.

[2] S. Alqahtani and et al, "Prediction of prostate cancer Gleason score upgrading from biopsy to radical prostatectomy using pre-biopsy multiparametric MRI PIRADS scoring system," *Scientific reports 10.1* (2020): 1-9.

[3] P. Banerjee, "A Guide on XGBoost hyperparameters tuning," *Kaggle*. (2020)

[4] T. Barrett and et al, "Ratio of Tumor to Normal Prostate Tissue Apparent Diffusion Coefficient as a Method for Quantifying DWI of the Prostate," *American Journal of Roentgenology* vol 205, (2015): 585-593. Doi: 10.2214/AJR.15.14338.

[5] J. C. Barsce, J. A. Palombarini and E. C. Martínez, "Automatic tuning of hyper-parameters of reinforcement learning algorithms using Bayesian optimization with behavioral cloning," *CoRR* vol abs/2112.08094, (2021)

[6] J. F. F. Bray, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians 68.6* (2018): 394-424.

[7] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," *Advances in neural information processing systems* vol. 24 (2011): 394-424.

[8] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins and D. D. Cox, "Hyperopt: a python library for model selection and hyperparameter optimization," *Computational Science & Discovery, 8(1)* p.014008 (2015).

[9] W. Chang and et al, "Prediction of Hypertension Outcomes Based on Gain Sequence Forward Tabu Search Feature Selection and XGBoost," *Diagnostics*, 11(5) (2021) p.792

[10] A. Chatterjee and et al, "Performance of ultrafast DCE-MRI for diagnosis of prostate cancer," *Academic radiology, 25(3)* (2018): 349-358.

[11] T. Chen, and C.Guestrin, "XGBoost: A Scalable Tree Boosting System," *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM (2016): 785–794

[12] F. Chollet, and et al, "Keras," *GitHub*. (2015).

[13] R. Cuocolo and et al, "Machine learning applications in prostate cancer magnetic resonance imaging," *European radiology experimental*. 3(1), (2019):1-8.

[14] C. De Vente, P. Vos, M. Hosseinzadeh, J. Pluim, and M. Veta, "Deep learning regression for prostate cancer detection and grading in bi-parametric MRI," *IEEE Transactions on Biomedical Engineering*. 68(2), (2020):374-383.

[15] R. Fraiman and G. Muniz, "Trimmed means for functional data," *Test 10*. (2001): 419–440. doi:10.1007/BF02595706.

[16] G. N. Dimitrakopoulos, A. G. Vrahatis, V. Plagianakos, and K. Sgarbas, "Pathway analysis using XGBoost classification in Biomedical Data," *In Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (2018): 1-6

[17] J. H. Hayes, and M. J. Barry, "Screening for prostate cancer with the prostate-specific antigen test: a review of current evidence," *Jama* 311(11) (2014): pp.1143-1149.

[18] A. M. Hotker and et al, "Assessment of Prostate Cancer Aggressiveness by Use of the Combination of Quantitative DWI and Dynamic Contrast-Enhanced MRI. AJR," *American journal of roentgenology* vol. 206,4 (2016): 756-63. doi:10.2214/AJR.15.14912.

[19] J. Jucevičius and et al, "Automated 2D Segmentation of Prostate in T2-weighted MRI Scans," *International journal of computers communication & control*, [S.l.], v. 12, n. 1, (2016): 53-60. ISSN 1841-9844.

[20] B. Liu and et al, "Prediction of prostate cancer aggressiveness with a combination of radiomics and machine learning-based analysis of dynamic contrast-enhanced MRI," *Clinical radiology*, 74(11) (2019): 896-e1.

[21] J. Liu and et al, "Predicting mortality of patients with acute kidney injury in the ICU using XGBoost model," *Plos one*. (2021);16(2):e0246306.

[22] S. López-Pintado and J. Romo, "On the Concept of Depth for Functional Data" *Journal of the American Statistical Association* v. 104, n. 486 (2009): 718-734. doi:10.1198/jasa.2009.0108.

[23] R. N. Low, D. B. Fuller, and N. Muradyan, "Dynamic gadolinium-enhanced perfusion MRI of prostate cancer: assessment of response to hypofractionated robotic stereotactic body radiation therapy," *American Journal of Roentgenology 197.4* (2011): 907-915.

[24] T. O'Malley and et al, "Keras Tuner," https://github.com/keras-team/keras-tuner (2019).

[25] F. Pedregosa and et al, "Scikit-learn: Machine learning in Python," *Journal of machine learning research* vol. 12, (2011): 2825-2830.

[26] I. Reda and et al, "Deep learning role in early diagnosis of prostate cancer," *Technology in cancer research & treatment*, 17, (2018) p.1533034618775530.

[27] A. Vaitulevičius and et al, "DCE MRI Modality Investigation for Cancerous Prostate Region Detection: Case Analysis ," unpublished.

[28] F. Wilcoxon, "Individual Comparisons by Ranking Methods," *Biometrics Bulletin* vol. 1, (1945): 80-83. doi:10.2307/3001968.

[29] C. J. Wu and et al, "DWI-associated entire-tumor histogram analysis for the differentiation of low-grade prostate cancer from intermediate-high-grade prostate cancer," *Abdom Imaging 40*, 3214-3221 (2015).