BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 69(3), 2021, Article number: e136749
DOI: 10.24425/bpasts.2021.136749

Multi-feature ensemble system in the renal tumour classification task

Aleksandra Maria OSOWSKA-KURCZAB[1*], Tomasz MARKIEWICZ[1,2], Miroslaw DZIEKIEWICZ[2], and Malgorzata LORENT[2]

[1] Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland
[2] Military Institute of Medicine, ul. Szaserów 128, 04-141 Warsaw, Poland

Abstract. Recently, the analysis of medical imaging is gaining substantial research interest, due to advancements in the computer vision field. Automation of medical image analysis can significantly improve the diagnosis process and lead to better prioritization of patients waiting for medical consultation. This research is dedicated to building a multi-feature ensemble model which associates two independent methods of image description: textural features and deep learning. Different algorithms of classification were applied to single-phase computed tomography images containing 8 subtypes of renal neoplastic lesions. The final ensemble includes a textural description combined with a support vector machine and various configurations of Convolutional Neural Networks. Results of experimental tests have proved that such a model can achieve 93.6% of weighted F1-score (tested in 10-fold cross validation mode). Improvement of performance of the best individual predictor totalled 3.5 percentage points.

Key words: medical imaging; renal cell carcinoma; convolutional neural networks; textural features; support vector machine; computer vision; deep learning.

1. Introduction

Recently, deep learning algorithms achieved outperforming results in almost any computer vision task. However, throughout the past decades, scientific society was also successfully working on specialised algorithms for feature description of images, especially, in specific tasks of medical image recognition. Combining both approaches might significantly increase the robustness of a created system, due to independence in image pattern characterisation. This research is aimed to investigate the possibility of creating a robust system of renal cell carcinoma (RCC) classification through ensemble learning by applying both approaches: deep learning and classical textural analysis.

Renal cancer is the 7th most common neoplastic disease in the UK according to the UK Cancer Research [1]. Diagnosis is hindered due to non-specific symptoms or even symptomless growing of tumour mass in the early stages of this disease. Nonetheless, early diagnosis and screening translate to extended survival rates and opportunities to apply the non-invasive treatment. A new approach is therefore needed for automation of the medical imaging examination.

Our research aims at finding a solution for this challenging problem through multi-feature classifiers, assembled in an ensemble with weighted voting. The system will be tested on the data set consisting of 8 types of renal neoplastic lesions. Features are generated in two distinct manners: textural and deep learning methods. Texture features are derived from raw grey-scale images, specifically from some statistical or structural patterns located in pixels co-occurrence. Those features are then passed to the Support Vector Machine (SVM) classifier. The second approach utilizes the transfer learning of pretrained ImageNet models. Both approaches separately demonstrated good overall results reaching 80–90% of weighted F1-score. However, their combination is able to yield higher overall performance. Results of the developed ensemble system were discussed taking into account the quality of the available data set and classification metrics.

2. Related Works

The topic of renal cell classification based on Computed Tomography (CT) scans has rarely been studied. Due to the lack of publicly open benchmark data sets, the results of various research papers cannot be objectively compared. However, due to the rapid development of deep learning for medical purposes, the subject of differencing renal cancer subtypes becomes more and more popular among scientific society.

Differentiation between a malignant and benign form of renal neoplasm is quite popular in the literature. This binary task can be solved through cross-training performed on retrained Inception-v3 model [2]. Simple transfer learning of Inception architecture applied to recognition between malignant clear cell carcinoma and benign oncocytoma on the basis of multiphasic CT has yielded sensitivity around 80% [3].

Only a few studies have shown results for the renal lesion subtypes differentiation. Convolutional Neural Networks (CNN) applied to histopathological images were able to facilitate detection of 3 types of renal carcinoma: chromophobe,

*e-mail: olaosows@gmail.com

Manuscript submitted 2020-10-08, revised 2020-10-08, initially accepted for publication 2020-10-24, published in June 2021

papillary and clear cell carcinomas [4] with results reaching 94% of accuracy. For comparison, GoogleNet achieved 85% accuracy in the same application based on 3-phase CT scans [5].

Morphological, textural and wavelet-based methods were used in 4-class renal neoplasm differentiation basing on histopathological images [6]. Basic features, such as contrast, correlation, energy, homogeneity and entropy, derived from Gray-Level Co-occurrence Matrix (GLCM), feeding the Bayesian classifier reached satisfactory performance around 90% of accuracy. Textural features have also found multiple applications in Computed Tomography Texture Analysis (CTTA) [7, 8]. Biomarkers for tumour differentiation were obtained from simple statistical features of raw image pixel intensities, such as average intensity, entropy or skewness.

Deep learning techniques are commonly utilized in lung cancer detection tasks since a higher incidence rate of this disease encouraged us to create larger data sets. Deep Belief Networks, Autoencoders [9] and 3D-CNNs [10, 11] proved to achieve state-of-the-art results in this domain.

Scientific society was not only addressing classification task but also staging through learnable image histogram-based deep neural networks [12] and segmentation with U-net [13, 14].

No study, to our knowledge, has considered the task of classification of 8 subtypes of renal tumours basing on single-phase CT scans. Moreover, no research was intended to combine the textural and deep learning methods in a unified ensemble system in order to increase robustness in multi-class prediction.

3. Method

This research is dedicated to analysing two separate methods of feature generation for renal neoplastic lesions differentiation. The first one facilitates textural methods, which gained wide popularity in medical imaging processing [15]. Three different algorithms had been tested and all of their outputs were passed to the SVM classifier. The second approach utilizes the transfer learning technique applied to pretrained deep ImageNet models.

Previous works on this topic have confirmed that both approaches enable to achieve performance metrics varying from 80–89% of F1-score for each of the aforementioned methods in 8-class classification problem [16, 17]. Influence of the data generation step was discussed, however, a combination of both methods was not considered.

3.1. Textural methods. Texture analysis focuses on the characterisation of image patterns on the basis of structural or statistical properties of raw pixel intensities. It usually performs additional preprocessing steps such as transformation into latent spaces to increase the readability of input data. Generally, image properties such as roughness or smoothness can be described by numerical features derived from pixel values with assumed neighbourhood pattern.

3.1.1. GLCM features. The first discussed algorithm – Gray-Level Co-occurrence Matrix (GLCM) [18] was initially proposed in 1973. Nonetheless, it still finds numerous applications in medical imaging analysis. Numerical features are computed from the co-occurrence matrix $h_{d\theta}$. Its values represent the number of co-occurrences of certain intensity values. Various hyperparameters of the symmetrical offset definition in GLCM were tested and 14 statistical measures were implemented as the input attributes to the classifier. Distances in this method were specified as $d \in \{1, 2, 3\}$, angles as $\theta = {}^{\pi}/_{4} \cdot n$ for $n \in \{0, 1, 2, \ldots 7\}$ and such definitions were used in numerical experiments. The resulting 3D matrices were vectorized either by simple average or by concatenation. Two ensemble members, derived from the GLCM method, were proposed.

3.1.2. SFTA features. Segmentation-based Fractal Texture Analysis (SFTA) [19] is a representation of fractal algorithms [20], which are usually applied to patterns with high local complexity. The authors of the algorithm proposed a two-stage process of features generation. The first step is the decomposition of the input image into sets of binary images through Two-Threshold Binary Decomposition. The final feature vector is prepared in the second step by computing the fractal dimensions of the region borders. The initial length of the feature vector, which is the main hyperparameter of SFTA, is denoted with n and is subject to changes. Three proposed ensemble members, based on this method, were generated with $n \in \{5, 6, 7\}$.

3.1.3. Unser features. Unser features [21, 22] were originally designed to solve the segmentation task. However, the universal character of the algorithm turned out to be also suitable to solve many other problems, for instance, classification. The principal operation performed in this algorithm is the application of a linear filter to shifted image segments. On the basis of such transformation, the classical descriptors can be computed, e.g. energy or correlation. Shift step s and filter size m are the only parameters in this method. In this research, single ensemble member was defined with $s = 5$ and $m = 3$.

Feature vectors generated by all these methods were passed to SVM classifier [23] with hyperparameters: Gaussian kernel (RBF), $C = 1000$ and $\gamma = 1$.

3.2. Deep learning models. Rapid development in computer vision owes to the implementation of Convolutional Neural Networks (CNNs) [24–26] that automatise the process of feature generation and simultaneously perform final prediction (e.g. regression or classification). However, the training of such architectures demands very large data sets and high computational power, which becomes a major bottleneck in the development of deep models.

Fortunately, these disadvantages can be solved with Transfer Learning. Instead of comprehensive training from scratch, a model trained on task A is adapted to new problem B. When domains of A and B are alike, the weights of initial layers can be frozen, which significantly reduces training time. On the whole, transfer learning can improve the performance of developed models and at the same time mitigate the problem of the data set's size requirement.

In this research, transfer learning will be applied using five different architectures, which won ImageNet Large Scale Visual

Recognition Challenge in the past few years: AlexNet [27], ResNet-18, ResNet-50 [28], Inception-v3 [29, 30] and Inception-ResNet-v2 [31].

Preliminary research has shown that deep learning models obtain the best results when frozen layers are in the initial parts of the networks. Adaptation of weights is not only done to the fully connected layers responsible for the final classification but also to weights of latter convolutional filters. Training parameters were set as follows: Adam solver, learning rate scheduler with initial value 1e-4, size of output layer was equal to the number of classes (here 8).

Each of the presented architectures formed an ensemble integrated with majority voting. Every ensemble consisted of 5 to 10 separate predictors to mitigate the problem of getting stuck in local minima. Moreover, each member was trained using the randomly selected learning data. Two distinct mechanisms have provided the independence of ensemble members: randomized division of training-validation data sets and diversity of size of fully connected layer appended to the pretrained part of the network (in AlexNet only).

3.3. Methodology. The data set was split into 10-fold cross-validation bins and each proposed model, textural and deep, was tested in a leave-one-out manner. The performance metrics applied in this paper were computed as an average of 10 runs of cross-validation. In later discussion, weighted F1-score [32] was selected as the most meaningful quality measure.

Key experiments conducted in this research consisted of 2 major steps. Firstly, all models were trained and separately evaluated. Secondly, the best combination of potential ensemble members was selected through an extensive search. Predictions of chosen models were integrated with weighted voting. As a result, a multi-feature ensemble model was built, and subsequently, its performance was evaluated. The summary of the proposed architecture is presented in Fig. 1.

4. Materials

The multi-feature ensemble model was evaluated on a database consisting of single-phase CT scans of 143 patients with 8 subtypes of renal neoplastic lesions [33, 34]. Raw images, results of histopathological examination and region contours were prepared in the laboratory of the Department of Pathology at the Military Institute of Medicine (Warsaw, Poland). Detailed information about data set distribution can be found in Table 1. Among 8 subtypes, there are five malignant subtypes (C, J, M, P, U), two benign (A, O) and a renal cyst (T). Although cyst is not considered as cancer itself, it might be sometimes misdiagnosed with other cancer types.

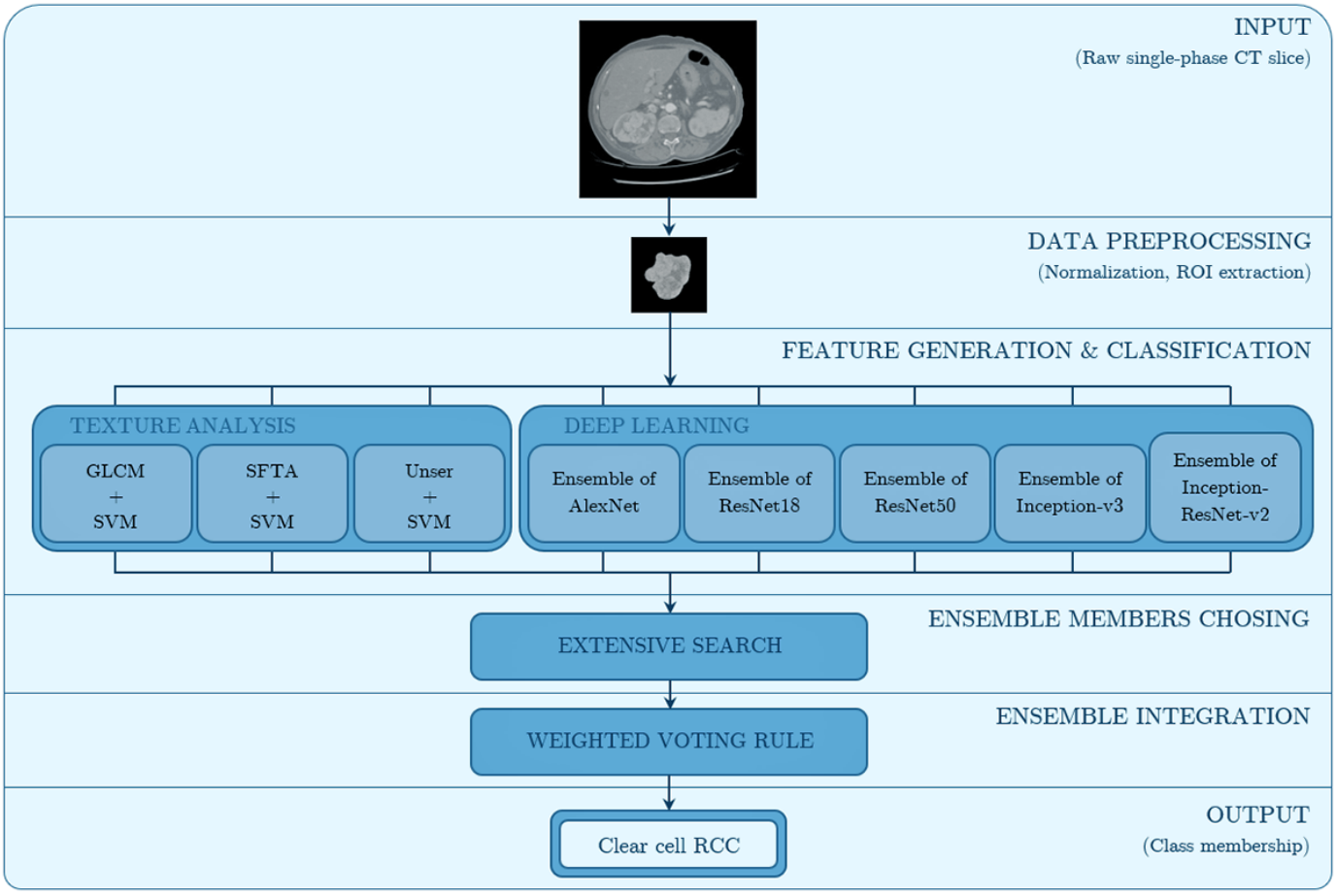

Fig. 1. General architecture of the classification system

Table 1
Renal lesion subtypes – full name, abbreviation and quantities

Renal neoplastic lesion (full name)	Number of scan frames (no. of patients)	Class label
Angiomyolipoma	97 (8)	A
Chromophobe Renal Cell Carcinoma	253 (20)	C
Clear Cell Renal Cell Carcinoma	692 (40)	J
Multilocular Cystic Renal Cell Carcinoma	164 (10)	M
Oncocytoma	108 (14)	O
Papillary Renal Cell Carcinoma	236 (26)	P
Urothelial Carcinoma	292 (11)	R
Renal Cyst	460 (14)	T

The preliminary experiments have shown that the generation method of the data set has a significant influence on the overall performance of the system. The fundamental assumption of the classification model is that segmentation has been already done. Hence, the position and contour of the neoplastic lesion are already known. Data preprocessing is divided into 3 main phases:

- **Normalisation.** Dicom images have to be transformed to unsigned integers of 8 bits (uint8) to match the common format used, e.g. in ImageNet models.
- **Region of interest (ROI) generation.** ROI represents the most essential part of a full CT frame and becomes an input to the model. Based on previous research [16, 17], one type of ROI generation was chosen. 100×100 mm region located in contour centre mass is cropped from the raw CT frame. The surrounding tissues are not included in the final image. Examples representing each class are depicted in Fig. 2.
- **Augmentation.** Due to a significant disproportion in the number of training examples, which is a result of the diversity of cancer incidence rates, augmentation of the data set has to be performed to balance the population of classes. Only rotation and cropping were used to create additional examples of images in order to prevent major distortions of image patterns.

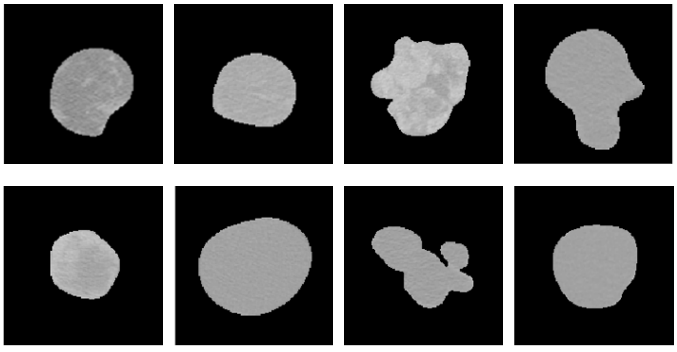

Fig. 2. Example frames (data inputs) from the preprocessed data set (from the upper left corner: A, C, J, M, and below: O, P, R, T)

The prepared data set is believed to be representative of the problem discussed in this study. However, few factors should be taken into consideration when the results of experiments are discussed. Firstly, the original data set is strongly unbalanced, images are rather small and of poor contrast. Therefore, patterns located in raw pixel intensities are rather fuzzy. Secondly, there is wide inter-patient diversity within each class. Finally, the model response is based on single frames, whereas medical specialists usually have access to the full series of scans in manually prepared diagnosis.

5. Results and discussion

Summary of proposed individual models and their parameters are presented in Table 2. For deep learning models, two major parameters are included. The first one is the level (name and type) of the cut-off layer, which is the last frozen layer in the model architecture that was subject to transfer learning. The second one is the number of ensemble members applied within this particular network. For instance, *res2a_branch2a* is the last frozen layer in ResNet18 and 5 models were trained in parallel.

Table 2
Parameters of independent models not associated in an ensemble

Model name	Parameters	
GLCM$^{(1)}$	aggregation: concatenation, $d \in \{1, 2\}$	
GLCM$^{(2)}$	aggregation: concatenation, $d \in \{1, 2, 3\}$	
SFTA$^{(1)}$	$n = 5$	
SFTA$^{(2)}$	$n = 6$	
SFTA$^{(3)}$	$n = 7$	
Unser	$s = 5, m = 3$	
Model name	**Cut-off layer**	**Ensemble size**
AlexNet$^{(1)}$	*fc6*	10
AlexNet$^{(2)}$	*fc7*	10
AlexNet$^{(3)}$	*fc8*	5
ResNet18	*res2a_branch2a*	5
ResNet50	*bn3b_branch2b*	5
Inception-v3	*conv2d_4*	5
Inception-ResNet-v2	*conv2d_4*	5

Initial results of individual members of the multi-feature ensemble, without integrating their answers, are presented in Table 3. Each entry corresponds to the independent model out-

Table 3
Results of independent models not associated in an ensemble. Presented values are an average between classes and 10-fold cross validation [in %]

Model name	Acc.	Prec.	Rec.	F1
GLCM (1)	81.63	82.01	81.63	81.59
GLCM (2)	82.55	82.97	82.55	82.53
SFTA (1)	88.58	89.09	88.58	88.64
SFTA (2)	88.32	88.82	88.32	88.37
SFTA (3)	88.35	89.14	88.35	88.44
Unser	77.02	77.78	77.02	77.05
AlexNet (1)	77.33	78.15	77.33	77.39
AlexNet (2)	80.44	80.85	80.44	80.36
AlexNet (3)	77.76	78.42	77.76	77.78
ResNet18	84.60	85.01	84.60	84.57
ResNet50	85.02	85.77	85.02	85.12
Inception-v3	89.46	89.83	89.46	89.47
Inception-ResNet-v2	90.06	90.45	90.06	90.10

come. Performance metrics (accuracy, precision, recall and F1) are included beside the abbreviated name of the model. All numerical values in the table represent an average of results of all classes which were obtained in cross-validation mode.

Results of independent models are varying from 77% to 90% of F1-score. Unser features combined with SVM and ensemble of 10 independently trained AlexNet networks are giving comparatively worse results, whereas ensemble of 5 Inception-ResNet-v2 units is reaching 90.1 ± 1.44% of F1-score in 10-fold cross-validation mode. Ensemble technique applied individually to Inception-ResNet-v2 eliminates fluctuations of performance, resulting from a random choice of training and testing data. Thanks to this, the F1-score was increased by 5–6pp on average. Although quality measures of performance are slightly improved, it should be emphasised that consequently, the computation time is several times higher.

The next step was an extensive search between members to obtain the best possible performance. The best-performing combination of models are marked in Table 2 and 3 with light blue background. Almost all types of feature generation methods took part in the final ensemble model, except ensembles built on the basis of AlexNets. At first glance, the algorithms such as Unser features, achieved considerably worse results. However, its predictions are valuable when models are combined in the final global ensemble. All possible combinations of members were tested, however, differences of accuracy between most combinations of members were negligible, especially, when the number of chosen units exceeded 10 members. Detailed information about the influence of the number of chosen predictors on performance is presented in boxplot form in Fig. 3.

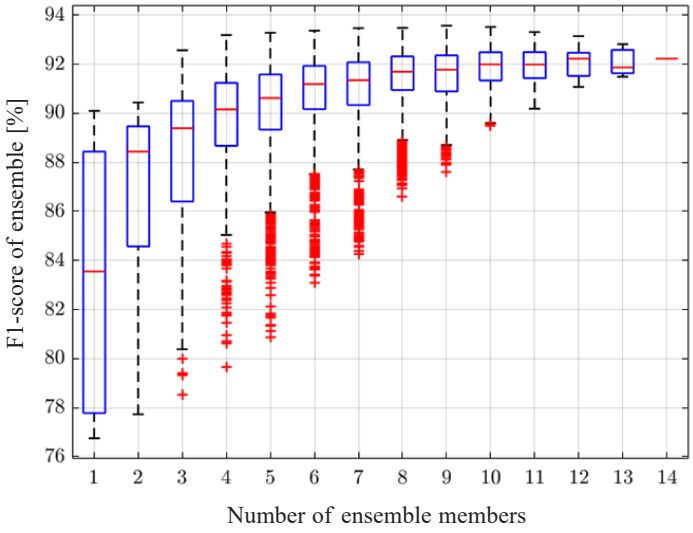

Fig. 3. Boxplot presenting influence of the number of chosen ensemble members on overall performance of the system (weighted F1-scores, in [%])

The results of the best combination of initially proposed members are presented in Table 4. Combining multiple methods of feature generation has significantly improved the classification metrics. The F1-score gains approximately 3.5pp with respect to the best individual predictor (an ensemble of 5 Inception-ResNet-v2 architectures). Final results of 93.66% should be considered satisfactory, especially when the complexity of the data set and the fuzziness of patterns are taken into account.

Table 4
Results of multi-feature ensemble model with the best-performing combination of models, average of 10-folds [in %, mean ± std (median)]

Accuracy	Precision	Recall	F1-score
93.64 ± 1.01 (93.82)	93.88 ± 0.96 (94.1)	93.64 ± 1.01 (93.82)	93.66 ± 1.01 (93.88)

Application of multi-feature ensemble technique with weighted majority voting has remarkably boosted the classification metrics when multiple independent techniques of feature generation are available. Various definitions of functions integrating predictions from independent classifiers were tested. All 4 quality metrics (accuracy, precision, recall, F1-score) were examined as weights in the weighted voting rule. Moreover, various normalization formulas (with L_p norm, where $p \in \{1, 2, \ldots 5\}$) were applied. The best results were obtained with F1 weights normalized using L_5 norm. However, the differences in system performance at the application of various weighting methods were negligible (below 0.1 percentage point).

Detailed information about system misclassifications can be found in the confusion matrix presented in Fig. 4. It was generated as a sum of confusion matrices for testing data in

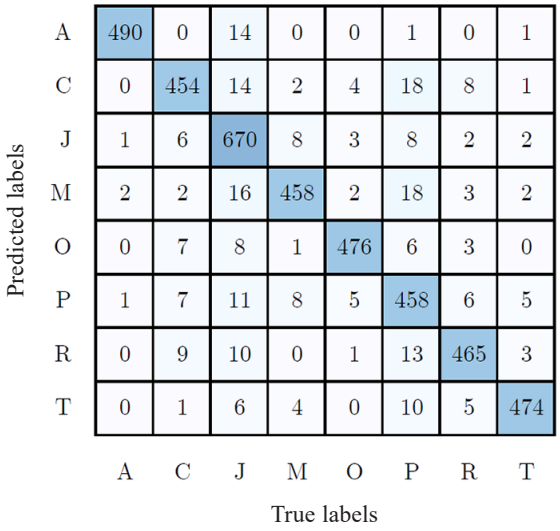

Fig. 4. Confusion matrix of multi-feature ensemble model, aggregated for 10 folds of cross validation

each fold of cross-validation. The majority of misclassifications have been committed in recognition of classes C-J-P and P-R-T, which reflects very well the human errors. Nonetheless, classes C, M and P are the most difficult in classification. F1 measure for these classes has reached 91.16%, 91.86% and 90.3%, respectively. On the contrary, the best results were obtained for classes A, O, T with F1-scores reaching 97.3%, 95.39% and 95.26%, respectively. Interesting results can be also observed in the heatmap of differences between confusion matrices (the best individual predictor vs. multi-feature ensemble) depicted in Fig. 5. The $(ij)^{th}$ element for $i \neq j$ of the matrix represents the difference between the number of errors committed by the best individual predictor (Inception-ResNet-v2) and a multi-feature ensemble. Therefore, the desired effect is negative elements outside the main diagonal and positive numbers on the main diagonal (representing exactly correct classifications). For the sake of visualisation, the improvements were coloured with green and deteriorations with the red colour. The differential matrix shows that the application of a multi-feature ensemble significantly eliminates most of the errors, especially in the classification of examples from classes M and P.

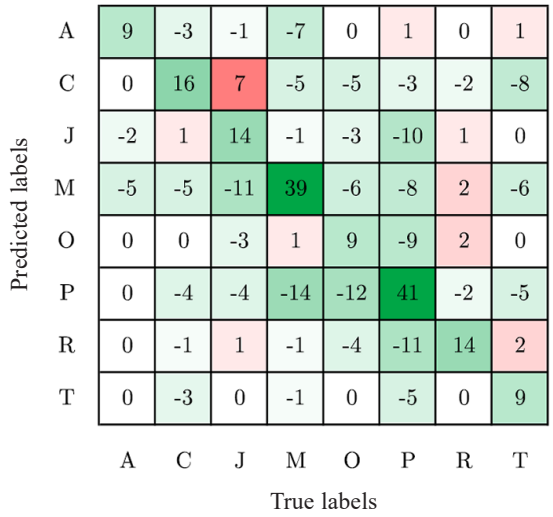

Fig. 5. Difference of confusion matrices between ensemble model and the best individual predictor (Inception-ResNet-v2). Colour represents direction of change: green – improvement, red – deterioration

Agreement of member predictions understood as a number of deciding votes, broken down into each class is presented on boxplot in Fig. 6. The shorter the interquartile range the more consistent distribution of ensemble member verdicts. The higher the level of the median, the more unanimous the final decision is. The maximum of the vertical axis is 9, which represents the total number of members in an ensemble. The most consistent answers of ensemble members are generated for classes A and O, for which only single answers have confidence lower than 9 and 7, respectively. Apparently, the ensemble is less confident in the prediction of classes with comparatively worse results in F1-score (classes C and P).

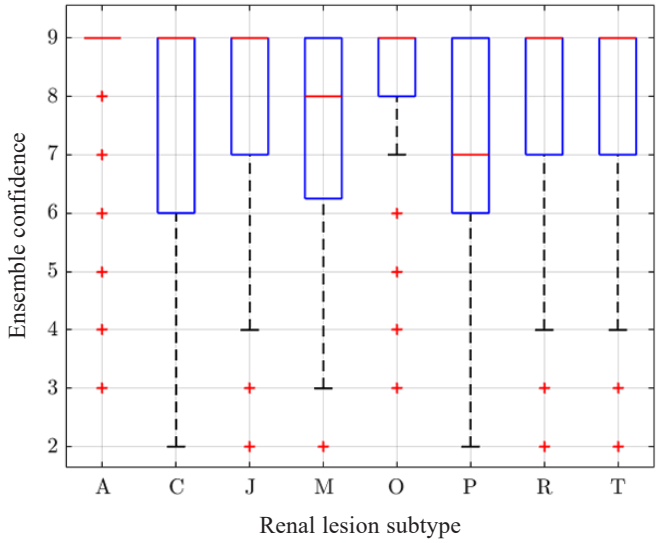

Fig. 6. Boxplot presenting confidence of ensemble predictions (confidence = number of deciding votes)

To summarize, the multi-feature ensemble model has shown high efficiency in the task of recognition of 8 subtypes of renal lesions. The combination of 2 families of independent approaches to feature generation eliminates the vast majority of misclassifications. Additional analysis is needed to target the reason for the remaining misclassified examples. Presumably, there is room for further improvement of the system by applying different data set generation methods. This line of research will be investigated in the nearest future. In conclusion, the system in the presented form will fulfil its role, though it makes the predictions based on the single frame extracted from the CT scan series. By applying the series of CT images it is possible to take into consideration the additional aspects, such as indications of the previous slide as well as the changing size of the neoplastic lesion. The smaller the evaluated object, the less confident the prediction is.

6. Conclusions and future work

The paper has proposed an automated computer system able to recognise 8 types of renal neoplastic lesions with high accuracy. The obtained results exceeded 93.6% of weighted F1-score on the testing data not taking part in training, evaluated in 10-fold cross validation mode. Generally, our results demonstrate a strong positive influence of the combination of independent types of features on overall system performance. The multi-feature ensemble model achieved the observed human ratio of mistakes. In addition, the incorrectly classified samples corresponded with the same types of data, where human experts made the errors the most often. Further analysis of misclassified examples may lead to new concepts in data preprocessing steps, especially regarding ROI extraction and data set augmentation.

The obtained results allow applying the system in regular hospital practice. Automation of hospital procedures might significantly reduce the waiting time for medical consultations, as well as decrease the number of renal cancer incidents which are diagnosed accidentally in imaging tests for some other diseases.

In the nearest future, few additional directions of research will be explored. Previous works indicated the substantial influence of the data set preparation stage on the final performance of the model. Therefore, new methods of ROI generation, especially the sliding window method will be examined. Furthermore, deblurring and super-resolution might effectively help in overcoming the problem of poor contrast and fuzziness of input images. Moreover, investigation of resizing techniques can improve the quality of data, since the small size of images is still a major bottleneck in obtaining better performance of the system. 3D-modelling and generative methods (such as Generative Adversarial Networks and Variational Autoencoders) applied to data set augmentation will also be investigated. Deployment of the model in everyday practice should be preceded by in-depth research on model explainability. The developed model should provide the medical specialist with insightful information about patient condition and attract attention to certain regions that would simplify and reduce the time of the diagnosis process.

Acknowledgements. This work was financially supported by the National Science Centre, Poland (grant no 2016/23/B/ST6/00621).

References

[1] "Kidney cancer statistics", Cancer Research UK, (2020). [Online]. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/kidney-cancer. [Accessed: 05-Oct-2020].

[2] L. Zhou et al., "A Deep Learning-Based Radiomics Model for Differentiating Benign and Malignant Renal Tumors", *Translational Oncology* 12(2), 292–300, (2019).

[3] H. Coy et al., "Deep learning and radiomics: the utility of Google TensorFlowTM Inception in classifying clear cell renal cell carcinoma and oncocytoma on multiphasic CT", *Abdominal Radiology* 44(6), 2009–2020, (2019).

[4] S. Tabibu, P.K. Vinod, and C.V. Jawahar, "Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning", *Scientific Reports* 9(1), 10509, (2019).

[5] S. Han, S.I. Hwang, and H.J. Lee, "The Classification of Renal Cancer in 3-Phase CT Images Using a Deep Learning Method", *Journal of Digital Imaging* 32, 638–643, (2019).

[6] Q. Chaudry, S.H. Raza, A.N. Young, and M.D.Wang, "Automated renal cell carcinoma subtype classification using morphological, textural and wavelets based features", *Journal of Signal Processing Systems* 55(1–3), 15–23, (2009).

[7] B. Kocak et al., "Textural differences between renal cell carcinoma subtypes: Machine learning-based quantitative computed tomography texture analysis with independent external validation", *European Journal of Radiology*, 107, 149–157, (2018).

[8] S.P. Raman, Y. Chen, J.L. Schroeder, P. Huang, and E.K. Fishman, "CT texture analysis of renal masses: pilot study using random forest classification for prediction of pathology", *Academic Radiology*, 12, 1587–1596, (2014).

[9] W. Sun, B. Zheng, and W. Qian, "Computer aided lung cancer diagnosis with deep learning algorithms", *Proceedings of the International Society for Optics and Photonics Conference* (2016).

[10] H. Polat and D.M. Homay, "Classification of Pulmonary CT Images by Using Hybrid 3D-Deep Convolutional Neural Network Architecture", *Applied Sciences*, 9(5), 940, (2019).

[11] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer Detection and Classification with 3D Convolutional Neural Network (3DCNN)", *International Journal of Advanced Computer Science and Applications (IJACSA)* 8(8), (2017).

[12] M.A. Hussain, G. Hamarneh, and R. Garbi, "Renal Cell Carcinoma Staging with Learnable Image Histogram-Based Deep Neural Network", *Lecture Notes in Computer Science*, 11861, 533–540, (2019).

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 9351, (2015).

[14] K. Yin et al., "Deep learning segmentation of kidneys with renal cell carcinoma", *Journal of Clinical Oncology* 37(15), (2019).

[15] J. Kurek et al., "Deep learning versus classical neural approach to mammogram recognition", *Bul. Pol. Acad. Sci. Tech. Sci.* 66(6), 831–840, (2018).

[16] A. Osowska-Kurczab, T. Markiewicz, M. Dziekiewicz and M. Lorent, "Textural and deep learning methods in recognition of renal cancer types based on CT images", *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, (2020).

[17] A. Osowska-Kurczab, T. Markiewicz, M. Dziekiewicz, and M. Lorent, "Combining texture analysis and deep learning in renal tumour classification task", *Proceedings of the Computational Problems of Electrical Engineering (CPEE)*, (2020).

[18] R.M. Haralick, K. Shanmugam, and I. Dinstein, "Textural Features for Image Classification", *IEEE Transactions on Systems, Man and Cybernetics*, SMC-3(6), (1973).

[19] A.F. Costa, G. Humpire-Mamani, and A.J.M. Traina, "An Efficient Algorithm for Fractal Analysis of Textures,", *Proceedings of 25th SIBGRAPI Conference on Graphics, Patterns and Images*, 39–46, (2012).

[20] P. Shanmugavadivu and V. Sivakumar, "Fractal Dimension Based Texture Analysis of Digital Images", *Procedia Engineering*, 38, 2981–2986, (2012).

[21] M. Unser, "Local Linear Transforms for Texture Analysis", *Proceedings of the 7th IEEE International Conference on Pattern Recognition (ICPR)*, II, 1206–1208, (1984).

[22] M. Unser, "Sum and difference histograms for texture classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 118–125, (1986).

[23] C. Cortes and V. Vapnik, "Support-vector network", *Machine Learning* 20(3), 273–297, (1995).

[24] Y. Bengio, "Learning Deep Architectures for AI", *Foundations and Trends in Machine Learning* 2 (1), 1–127, (2009).

[25] Y. Bengio, Y. LeCun, and G. Hinton, "Deep Learning", *Nature* 521, 436–444, (2015).

[26] I. Goodfellow, Y. Bengio, and A. Courville: Deep Learning, MIT Press, 2016.

[27] A. Krizhevsky, I. Sutskever, and G. Hinton, "Image net classification with deep convolutional neural networks", *Advances in Neural Information Processing Systems* 25, 1–9, (2012).

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", *Proceedings of the 29th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, (2016).

[29] C. Szegedy et al., "Going deeper with convolutions", *Proceedings of the 28th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9, (2015).

[30] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", *Proceedings of the 29th IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826, (2016).

[31] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inceptionv4, Inception-ResNet and the impact of residual connections on learning", *Proceedings of 31st Association for the Advancement of Artificial Intelligence on Artificial Intelligence (AAAI)*, 1–12, (2016).

[32] P.N. Tan, M. Steinbach, and V. Kumar: Introduction to Data Mining, Pearson Education, Boston, 2006.

[33] H. Moch, A.L. Cubilla, P.A. Humphrey, V.E. Reuter, and T.M. Ulbright, "The 2016 WHO Classification of Tumours of the Urinary System and Male Genital Organs-Part A: Renal, Penile, and Testicular Tumours", *European Urology*, 70(1), 93–105, (2016).

[34] T. Gudbjartsson et al., "Renal oncocytoma: a clinicopathological analysis of 45 consecutive cases", *BJU International* 96(9), 1275–1279, (2005).