DEEP RANDOM FORESTS FOR SMALL SAMPLE SIZE PREDICTION WITH MEDICAL IMAGING DATA

Alexander Katzmann[⋆‡] *Alexander Muehlberg*[⋆] *Michael Suehling*[⋆]
Dominik Nörenberg[†1] Julian Walter Holch[†2] Horst-Michael Gross[‡]

[⋆]Siemens Healthcare GmbH, Department CT R&D Image Analytics, 91301 Forchheim, Germany

[†]University Hospital Großhadern, Ludwig-Maximilians-University München
[1] Department of Radiology, [2]Department of Internal Medicine III, Comprehensive Cancer Center, Marchioninistrasse 15, 81377 Munich, Germany

[‡]Ilmenau, University of Technology, Neuroinformatics and Cognitive Robotics Lab, 98693 Ilmenau, Germany

ABSTRACT

Deep neural networks represent the state of the art for computer-aided medical imaging assessment, e.g. lesion detection, organ segmentation and disease classification. While for large datasets their superior performance is a clear argument, medical imaging data is often small and highly heterogeneous. In combination with the typical parameter amount in deep neural networks, this often leads to overfitting and results in a low level of generalization performance. We propose a straight-forward combination of random forests and deep neural networks for superior performance on medical imaging datasets with only small data, and provide an extensive evaluation of survival prediction for metastatic colorectal cancer patients using computed tomography imaging data, with our proposed method clearly outperforming other approaches.

***Index Terms*—** Ensemble learning, Random forests, Survival prediction

1. INTRODUCTION

Training of convolutional neural networks (CNNs) has become a widely employed technique for medical image classification [1, 2, 3]. Given large datasets, CNNs were shown to be significantly superior over other machine learning techniques for a variety of medical tasks, at times archieving an accuracy on par with gold standard human assessment [4, 5, 6, 7]. When only small data is available, their application becomes somewhat complicated, and especially for medical applications typically only small datasets are available.

In medical applications, the problem space is often underdetermined due to large problem spaces on the one hand, and a lack of large and publicly accessible medical datasets on the other, which may lead to overfitting. Although there exist techniques to tackle this issue, it typically results in a significant number of manual adjustments for regularization and data augmentation.

We propose a novel, deep-learning-based approach for small datasets using a model ensemble technique inspired by and constructed analogously to the well known random forest classifier, which can help to significantly reduce the need for interactive model augmentation and regularization while providing superior classification performance on unseen data.

2. BACKGROUND

Besides from deep neural networks, random forests (RFs) are one of the most commonly employed machine learning algorithms as they are easy-to-use and remarkably robust. Their applicability is not limited to typical classification tasks, but encompasses applications such as regression, survival analysis, and others. While deep neural networks (DNNs) automatically derive discriminative features from data through optimization, RFs require handcrafted features that are usually based on a-priori knowledge (e.g. Radiomics features [8]). Especially with medical data this becomes an important issue, as medical research is often explorative, intuition-driven, and has the explicit goal of identifying novel, e.g. visual, biomarkers. A combination of DNNs and RFs might therefore result in a solution which:

- handles medical imaging data with only small datasets,
- requires no prior knowledge-driven, hand-crafted feature design,
- outperforms deep neural networks and simple ensemble techniques in terms of generalization error in unseen data.

Int. Symposium on Biomedical Imaging (ISBI), Iowa City, USA, pp.1543-47, IEEE 2020

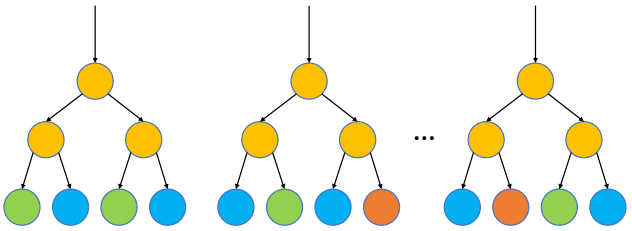

Fig. 1. Visualization of the architecture of Deep Random Forests. The model consists of several decision trees with DNNs as nodes (yellow), which split the data with the extracted semantic features. The final classification is done either by a leaf classifier (blue) or directly for pure nodes, i.e. nodes of only one class (green/red). Each tree is trained with a bootstrapped subset of the original data. All tree predictions are averaged for classification

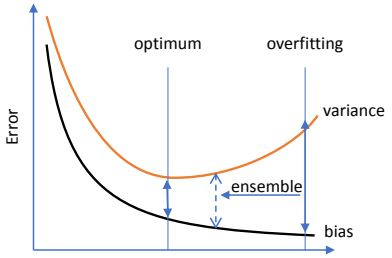

Fig. 2. Visualization of the effects on the bias-variance tradeoff. Combining multiple classifier results by averaging multiple low-bias/high-variance classifiers effectively approaches the point of optimal tradeoff.

We propose such a combination, using well-known prinicples, and demonstrate its applicability to medical image classification with the example of lesion-based one-year survival prediction in patients with metastatic colorectal cancer.

3. DEEP RANDOM FORESTS

While we give a basic introduction to the proposed method in Sec. 3.1, we describe the concrete training procedure in Sec. 3.2 and 3.3 and provide an experimental evaluation in Sec. 4, which is discussed in Sec. 5.

3.1. Random Forest Classifiers and Deep Ensembles

RFs were introduced in 2001 by Breiman [9], while their origins date back to 1995 [10] and earlier. Basically, RFs are an ensemble of decision trees where each tree is independently trained on a bootstrapped subset of the training data (i.e. samples drawn with replacement). Trees are correlated only in the way that they are trained with partially common training data, and the original publications explicitely emphasized the

Algorithm 1 Deep Random Forest Training

for K `trees` **do**
  Draw N samples X_k, Y_k from X, Y with replacement
  Train decision tree $DT(X_k) : X_k \rightarrow Y_k$:
    Train neural network to solve $X_k \rightarrow Y_k$
    if $>= 2$ `classes` & `depth` $\leq$ `max. depth` **then**
      Extract $M = |\Phi(X_k)|$ features
      Randomly select $m = \lfloor \sqrt{M} \rfloor$ features
      Select split feature and threshold T
      Train subtrees for both subsets (Step 3)
    else
      Create leaf

bootstrapping idea behind this methodology [9, 10]. As decision trees tend to overfit the data, RFs aggregate uncorrelated trees, resulting in an effective regularization, that shifts the low-bias/high-variance approximation of a single decision tree towards a low-variance approximation (see Fig. 2).

DNNs are known to generalize better when combined in model ensembles. Recent work on this topic, inside and outside the medical field, mentions various applications of, e.g. horizontal, vertical, and snapshot ensembling methods [11, 12, 13, 14]. Indeed, it has been shown that the widespread residual networks (ResNets) [15] behave comparably to an ensemble of relatively shallow networks [16]. Furthermore, random forest-like structures can be directly represented within neural networks (Fig. 3). However, with only few exceptions [17, 18], bootstrapping methods are largely unkown for deep learning (DL), although they have long been proposed for neural networks [19, 20]. While there was research on a direct combination of RFs and DL (most noteworthy *gcForests* [21]), previous work shifted away from the basis of neural network by directly employing random-forest-like structures as layers, and was inferior in comparison to current state-of-the-art approaches on some widely available benchmark datasets with more data [22, 23]. Our work, in contrast, is substantially different in the sense that we propose randomized decision tree ensembles, called *Deep Random Forests*, based on complete networks and not randomized layers, without introducing restrictions regarding the actual network architecture.

3.2. Model Architecture

RFs typically consist of decision trees of weak classifiers, that use one or a few features to determine a criterion-based (e.g. Gini coefficient, or entropy) split. We adopt the idea of only using a minor subset of the available features at once, since it acts as a regularization itself, but employ semantic features derived by deep neural networks for splitting. A visualization of the basic architecture can be found in Fig. 1. For the construction of Deep Random Forests, we propose Algorithm 1. While the training procedure still resembles RFs,

each node within Deep Random Forests contains a DNN that has been trained to solve the actual classification task for its particular subset. Finally, the last layer of the DNN is removed and the derived features are used (see Sec. 3.3). In analogy to RFs, each subnode is trained with those samples being most similar, and thus difficult to discriminate, with respect to a specific feature. Since the training set is different for each node, the network is obligated to always learn characteristics that are discriminative within the concrete subset. Formally, inter-cluster variance is used to discriminate easy-to-split samples, while intra-cluster variance is maximized to discriminate hard-to-split samples within each subset. To accelerate feature inference, subtree DNNs are initialized with the weights of their respective parent classifier, a technique widely used in neural architecture search (NAS) [24, 25]. For each subtree, the samples are reweighted according to their new class distributions.

3.3. DNN Splitting Criterion

As already mentioned in Sec. 3.2, first a network D with L layers is trained to solve the classification task $X_k \rightarrow Y_k$ at each node, with (X_k, Y_k) being the bootstrapped training set of tree k. The network D can be interpreted as a concatenation $D(X_k) = (h_L \circ h_{L-1} \circ \cdots \circ h_1)(X_k)$ of L functions $h_i : 1 \leq i \leq L$, where the last layer h_L is typically the final output activation function. When training D to solve the classification task, the features at layer h_{L-1} are optimized to contain information that is discriminative for the concrete sample subset, so that this layer is used for the split creation.

For each sample, M features $\Phi(X_k) = (h_{L-1} \circ \cdots \circ h_1)(X_k)$ with $|\Phi| = M$ are extracted. Of these, $m = \lfloor\sqrt{M}\rfloor$ features ϕ are drawn randomly and considered independently for the split. For each feature ϕ_j the optimal threshold T_j^* for partitioning the set (X_k, Y_k) is calculated by maximizing the split value V_j of the feature j. V_j is calculated by using the Gini coefficient based on the relative class probabilities p_i within both subsets as:

$$\sum_\theta \frac{\sum_{\alpha=1}^{n_\theta}\sum_{\beta=1}^{n_\theta}|p_\alpha - p_\beta|}{2n_\theta \sum_{\alpha=1}^{n_\theta} p_\alpha}$$
$$p \in P_\theta, \quad P_\theta \in \{P(Y_k|\phi_j < T_j), P(Y_k|\phi_j \geq T_j)\} \tag{1}$$

with relative class probability distributions P_θ for n_θ classes in subsets θ and features ϕ_j lower than, or greater than or equal to the threshold T_j, respectively. For each subset, a derived node is trained with the same procedure until the subset is either pure (i.e. consists of samples of only one class; red and green nodes in Fig. 1), or until a specified maximum depth is reached. In the latter case, the classifier output is used for prediction (blue nodes, Fig. 1).

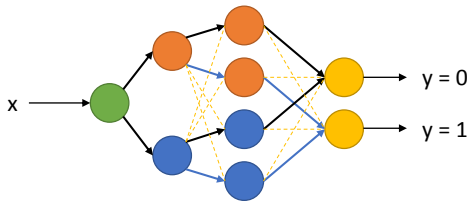

Fig. 3. Random forest-like structures within neural networks. Positive (black), nearly zero (yellow) and negative (blue) weights are represented by bold (high-magnitude), or dashed lines (low-magnitude). The forests consists of two trees of depth 2 (orange and blue neurons). The output layer consists of a weighted average of the tree outputs, which effectively implements a (non-randomized) forest classifier.

4. EXPERIMENTS

We have examined our model at different granularities to demonstrate the impact of all components. As baseline, we employed a RF classifier of 1,000 trees trained with a) the maximum lesion diameter (RECIST [26]), and b) the Radiomics signature from [8]. Additionally, we trained c) a vanilla convolutional neural network, d) a complex neural network based on ResNet [15], e) a deep decision tree as proposed using ResNets as node classifiers (Deep DT), and f) a Deep Random Forest. All networks were built with 3 blocks (normal/residual convolution), followed by a global average pooling and a softmax output. The blocks were built with 3x3 convolutions, batch normalization and leaky ReLU activation, followed by either max pooling (ConvNet) or strided convolutions (ResNet). Residual blocks were built from 5 convolutional blocks with residual connections. The maximum depth of deep trees was set at 3, ensembles each contained 30 elements, i.e. networks or trees. All metrics were calculated using 10 times 10-fold grouped cross validation, i.e. 100 classifiers per task, with identically varied random seeds using micro-averaging. Confidence intervals were calculated using bootstrapping [27] until convergence ($\epsilon < 10^{-3}$), significance is reported with $\alpha = .05$.

4.1. Metastatic Colorectal Cancer

We used the dataset of [28] to predict the one-year survival of patients with metastatic colorectal cancer (mCRC) from single liver metastases computed tomography images (1,282 longitudinal annotations, 885 pos., 397 neg., from 491 lesions and 104 patients). As in the original publication, all lesions were masked and the slice with the largest RECIST diameter was used to predict one-year survival based on a 64x64 image of 80x80mm in world coordinates. Since longitudinal data are not necessarily available in practice, *only one timepoint* was used for prediction to prevent extrapolation.

The results in Tab. 1 correspond to the above analysis.

Table 1. Results on mCRC dataset, 95% CIs

	Sensitivity	Specificity
Radiomics	.563 [.532,.595]	.509 [.455,.575]
RECIST	.584 [.557,.613]	.535 [.480,.591]
ConvNet	.537 [.507,.567]	.476 [.427,.527]
ResNet	.549 [.519,.582]	.532 [.476,.583]
Deep DT	.560 [.521,.598]	.534 [.469,.600]
Deep Forest	**.596** [.566,.625]	**.555** [.500,.615]
	ϕ-coefficient	AUC
Radiomics	.063 [.003,.122]	.556 [.518,.591]
RECIST	.104 [.048,.157]	.562 [.527,.594]
ConvNet	.012 [-.037,.060]	.524 [.492,.555]
ResNet	.071 [.023,.128]	.570 [.538,.605]
Deep DT	.082 [.018,.144]	.577 [.534,.618]
Deep Forest	**.132** [.075,.189]	**.610** [.574,.647]

ResNet outperforms the ConvNet approach, which might be due to the ensemble-like behavior of ResNet ([16]; see Sec. 3.1). However, due to the problem complexity and the small amount of data, ResNet provides only mediocre results ($\phi = .071$, $AUC = 57.0\%$), and is comparable to the results of RECIST-diameter- or Radiomics-signature-based prediction. The deep decision tree provides a slightly higher AUC (57.7%), although the difference is not significant. The Deep Random Forest outperforms the other approaches in each metric with a ϕ-coefficient of 13.2%, which is significantly higher than the Radiomics ($t(102) = 2.35, p = .021$, two-tailed), ConvNet ($t = 4.5, p = 1.8 \cdot 10^{-5}$) and ResNet ($t = 2.21, p = .029$) approaches, and an area under the curve of 61.0%, again significantly higher than the RECIST, Radiomics, ConvNet and ResNet approaches ($t(102) = 2.72/2.93/5.00/2.26, p = .008/.004/2.4 \cdot 10^{-6}/.026$).

5. DISCUSSION

As indicated in Sec. 4, the application of standard deep learning architectures when having only small datasets is not always possible. Without requiring prior knowledge or handcrafted features, as opposed to classical Radiomics-based approaches, Deep Random Forests outperformed all other tested classifiers in terms of all tested metrics on the given dataset. It is noteworthy that classifying survival based on one single lesion is a rather difficult task. However, given the prevalence of colorectal cancer it can be assumed that deriving even only few additional information can lead to a significantly improved patient wellbeing on the large scale and can therefore be of high clinical value.

Deep Random Forests could provide a highly beneficial, easy-to-use framework for medical image classification, where data is usually sparse and difficult to acquire. Particularly explorative research could benefit from our approach, as it highlights small differences when only a small amount of data is available. Future work should analyze whether the method improves results for large medical image datasets, too.

We plan to analyze the applicability of other splitting criteria, as well as ways to improve the computational effiency. The training was done on an HPC platform and could be done within one week for all runs. Training a single forest using state-of-the-art hardware takes about one day with an RTX 2080 Ti.

While this paper provides an example of a classification task, future work should cover the applicability for other purposes, e.g. medical image segmentation. Variants of RFs have been used for a variety of other uses, with survival regression being of particular interest for medicine. While this topic can not be covered within this paper, an extension of the approach to survival regression could be an interesting topic for future research.

Acknowledgements

This work has received funding from the German Federal Ministry of Education and Research as part of the PANTHER project under grant agreement no. 13GW0163A.

6. REFERENCES

[1] Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim, "Deep learning applications in medical image analysis," *Ieee Access*, vol. 6, pp. 9375–9389, 2017.

[2] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

[3] Andreas Maier, Christopher Syben, Tobias Lasser, and Christian Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, 2019.

[4] Anton S Becker, Michael Mueller, Elina Stoffel, Magda Marcon, Soleen Ghafoor, and Andreas Boss, "Classification of breast cancer in ultrasound imaging using a generic deep learning analysis software: a pilot study," *The British journal of radiology*, vol. 91, no. xxxx, pp. 20170576, 2018.

[5] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke

Hermsen, Quirine F Manson, Maschenka Balkenhol, et al., "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[6] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, et al., "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *European Journal of Cancer*, vol. 113, pp. 47–54, 2019.

[7] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115, 2017.

[8] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, René Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature communications*, vol. 5, pp. 4006, 2014.

[9] Leo Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] Tin Kam Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*. IEEE, 1995, vol. 1, pp. 278–282.

[11] Noel CF Codella, Q-B Nguyen, Sharath Pankanti, DA Gutman, Brian Helba, AC Halpern, and John R Smith, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM Journal of Research and Development*, vol. 61, no. 4/5, pp. 5–1, 2017.

[12] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger, "Snapshot ensembles: Train 1, get m for free," *arXiv preprint arXiv:1704.00109*, 2017.

[13] Ashnil Kumar, Jinman Kim, David Lyndon, Michael Fulham, and Dagan Feng, "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE journal of biomedical and health informatics*, vol. 21, no. 1, pp. 31–40, 2016.

[14] Jingjing Xie, Bing Xu, and Zhang Chuang, "Horizontal and vertical ensemble with deep representation for classification," *arXiv preprint arXiv:1306.2759*, 2013.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[16] Andreas Veit, Michael J Wilber, and Serge Belongie, "Residual networks behave like ensembles of relatively shallow networks," in *Advances in neural information processing systems*, 2016, pp. 550–558.

[17] Jürgen Franke and Michael H Neumann, "Bootstrapping neural networks," *Neural computation*, vol. 12, no. 8, pp. 1929–1949, 2000.

[18] Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich, "Training deep neural networks on noisy labels with bootstrapping," *arXiv preprint arXiv:1412.6596*, 2014.

[19] Zhi-Hua Zhou, Yuan Jiang, Yu-Bin Yang, and Shi-Fu Chen, "Lung cancer cell identification based on artificial neural network ensembles," *Artificial Intelligence in Medicine*, vol. 24, no. 1, pp. 25–36, 2002.

[20] Zhi-Hua Zhou, Jianxin Wu, and Wei Tang, "Ensembling neural networks: many could be better than all," *Artificial intelligence*, vol. 137, no. 1-2, pp. 239–263, 2002.

[21] Zhi-Hua Zhou and Ji Feng, "Deep forest," *arXiv preprint arXiv:1702.08835*, 2017.

[22] Kevin Miller, Chris Hettinger, Jeffrey Humpherys, Tyler Jarvis, and David Kartchner, "Forward thinking: Building deep random forests," *arXiv preprint arXiv:1705.07366*, 2017.

[23] Lev V Utkin and Mikhail A Ryabinin, "A siamese deep forest," *Knowledge-Based Systems*, vol. 139, pp. 13–22, 2018.

[24] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter, "Neural architecture search: A survey," *arXiv preprint arXiv:1808.05377*, 2018.

[25] Esteban Real, Sherry Moore, Andrew Selle, Saurabh Saxena, Yutaka Leon Suematsu, Jie Tan, Quoc V Le, and Alexey Kurakin, "Large-scale evolution of image classifiers," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2902–2911.

[26] Elizabeth A Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H Schwartz, D Sargent, Robert Ford, Janet Dancey, S Arbuck, Steve Gwyther, Margaret Mooney, et al., "New response evaluation criteria in solid tumours: revised recist guideline (version 1.1)," *European journal of cancer*, vol. 45, no. 2, pp. 228–247, 2009.

[27] Bradley Efron and Robert Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statistical science*, pp. 54–75, 1986.

[28] Alexander Katzmann, Alexander Muehlberg, Michael Sühling, Dominik Noerenberg, Julian Walter Holch, Volker Heinemann, and Horst-Michael Groß, "Predicting lesion growth and patient survival in colorectal cancer patients using deep neural networks," 2018.