1 (2025) 100008

Contents lists available at ScienceDirect

EANM Innovation

journal homepage: www.sciencedirect.com/journal/eanm-innovation

Impact of regularization in optimizing distance-encoding biomorphic-informational neural networks for small nuclear medicine datasets

Boglarka Ecsedi [a,b], Amine Boukhari [a,c], Clemens P. Spielvogel [d,k], David Haberl [d], Zsombor Ritter [e], Ralph A. Bundschuh [f,g,h,i], Constantin Lapa [j], Marcus Hacker [d], Mathieu Hatt [c], Laszlo Papp [a,k,*]

[a] Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria
[b] Georgia Institute of Technology, Atlanta, GA, USA
[c] LaTIM, INSERM, UMR, 1101, Univ Brest, Brest, France
[d] Division of Nuclear Medicine, Medical University of Vienna, Vienna, Austria
[e] University of Pécs, Medical School, Department of Medical Imaging, Division of Nuclear Medicine, Pécs, Hungary
[f] Department of Nuclear Medicine, University Hospital Carl Gustav Carus Dresden, Dresden, Germany
[g] Institute of Radiopharmaceutical Cancer Research, Helmholtz Zentrum Dresden-Rossendorf (HZDR), Rossendorf, Germany
[h] German Cancer Consortium (DKTK), Partner Site Dresden, Dresden, Germany
[i] National Center for Tumor Diseases (NCT), NCT/UCC Dresden, a Partnership Between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, And Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Germany
[j] Nuclear Medicine, Medical Faculty, University of Augsburg, Augsburg, Germany
[k] Comprehensive Center for AI in Medicine, Medical University of Vienna, Vienna, Austria

ARTICLE INFO

Keywords:
Biomorphic computing
Neural network regularization
Distance encoding
Machine learning
Medical AI

ABSTRACT

Introduction: To date, small, imbalanced datasets are considered challenging to efficiently train deep learning (DL) models, especially in the medical domain. Consistently, most Artificial Intelligence (AI) approaches in conjunction with small datasets rely on shallow radiomics where traditional machine learning (ML) is utilized for analysing image-derived features. In this study, we evaluate a recently introduced spatial neural network scheme called Distance-Encoding Biomorphic-Informational Neural Network (DEBI-NN), which trains spatial coordinates of artificial neuron coordinates instead of weights, that are then calculated from neuron distances. This technique dramatically reduces the number of parameters to train, thereby making DEBI-NN eligible for the analysis of small, imbalanced datasets. We refer to this property as spatial plasticity. We hypothesized that DEBI-NNs could systematically outperform baseline NN models in small clinical datasets while requiring less regularizations to be implemented, as spatial plasticity may have self-regularization properties. To test our hypothesis, we aimed to compare DEBI-NNs with baseline NNs while relying on various regularization techniques to investigate how DEBI-NNs perform in the presence of regularizers in small multi-centric medical imaging datasets.

Methods: Three multi-centric datasets were collected including diffuse large B-cell lymphoma (DLBCL) [^{18}F]FDG positron emission tomography (PET)/computed tomography (CT) with clinical parameters to predict 2-years event-free survival; the head and neck [^{18}F]FDG PET/CT dataset from the 2022 MICCAI challenge (HECKTOR), predicting human papillomavirus status; and [^{68}Ga]Ga-PSMA-11 (PSMA-11) PET/CT as well as PSMA-11 PET/magnetic resonance imaging (MRI) cases to predict histopathology-provided International Society of Urological Pathology (ISUP) grades as low (ISUP ≤ 2) and high (ISUP > 2) risk. Per cohort, 5 different network configurations having 1, 2 and 3 hidden layers and neuron count configurations were defined. Per configuration, DEBI-NNs had 7 regularization techniques and baseline NNs had 6 regularization configurations, totalling $2^7 = 128$ and $2^6 = 64$ regularization variants per network scheme to train and evaluate. Test balanced accuracy (BACC) was measured for each model and correlation of the test BACC in the presence of regularization techniques was evaluated in DEBI-NN and baseline NN models.

* Corresponding author. Center for Medical Physics and Biomedical Engineering & Comprehensive Center for AI in Medicine Medical University of Vienna Waehringer Guertel 18-20 1090 Vienna, Austria.
E-mail address: Laszlo.papp@meduniwien.ac.at (L. Papp).

https://doi.org/10.1016/j.eanmi.2025.100008
Received 5 August 2025; Received in revised form 4 September 2025; Accepted 11 September 2025
Available online 12 September 2025

Results: The best-performing DEBI-NN models yielded 84.5 %, 80 % and 80.5 % BACC in DLBCL, HECKTOR and PROSTATE datasets, respectively. In contrast, the highest-performing baseline NN models yielded 71.9 %, 77.3 % and 77.3 % BACC in the same cohorts, respectively. In addition, baseline NNs required the implementation of more regularization techniques to increase test BACC from an average test BACC of 53 % (no regularization) to 60 % (6 regularizations), while DEBI-NNs needed no regularization to achieve 62 % BACC. In return, DEBI-NN BACC monotonously fell down to 56 % BACC as the number of regularizations increased.
Conclusions: DEBI-NNs exhibit a significantly simpler training complexity compared to baseline NNs, while they also outperform baseline NNs with the presence of minimal or no regularization techniques. Our results strongly imply that DEBI-NNs have a potential to pave the way for the utilization of neural networks in small and imbalanced medical datasets, which the field of medical imaging research routinely operates with.

1. Introduction

To date, small datasets are considered challenging to efficiently train deep learning (DL) models, especially in the medical domain [1,2]. Medical imaging datasets readily available for DL development tend to be small compared to other domains within computer vision. In addition, medical imaging datasets are often collected from multiple sites to ensure generalizability and to increase the overall sample size. This often leads to a large domain shift across various imaging centers, time, machines, acquisition protocols, or reconstruction settings, introducing additional complexity for machine learning (ML) models [3], making them difficult to yield consistent, reliable predictions in new unseen data. Furthermore, these datasets tend to be highly imbalanced which can result in ML algorithms failing to correctly detect rare cases of interest. Despite previous efforts to harmonize both medical imaging data and extracted features [4], it remains a challenge to successfully train efficient and generalizable DL models on small medical imaging datasets [1].

Since the early days of machine learning research, radiomics combined with traditional ML approaches have been one of the most commonly utilized techniques to process medical imaging data and to support clinical decision-making for their interpretability, low cost and ease of training compared to their deep learning counterparts. Radiomics involves semi-automatically extracting a large number of quantitative features from medical images which then can be analysed along with other clinically relevant data using traditional ML algorithms, including e.g., extreme gradient boosting (XGBoost), random forests (RF) or support vector machines (SVMs). Research indicates that relying solely on these models with small sample sizes often yields unreliable results and falls short in capturing complex patterns within the data [1].

Following the success of DL models in other domains, the clinical community has increasingly sought to leverage their representational power and hone their advantages for medical image analysis as well. The convolutional neural network (CNN) [5] was one of the first revolutionary model families in computer vision research which was then followed by the residual neural network and the attention-based vision transformer (ViT) [6,7], quickly making breakthroughs in classical computer vision tasks including classification, detection, and segmentation, starting to replace traditional methods in many fields and applications. However, DL methods typically require large, representative and balanced datasets to achieve optimal performance. In the absence of such high-quality, abundant data, performance deficiencies can be mitigated with techniques such as data augmentation, transfer learning and carefully-designed initialization [8], or customized loss functions. Scarce, imbalanced, yet highly complex datasets still impose a huge constraint to all of the newer DL approaches. Another concern is that DL methods often lack by-design explainability and require complex interpretability techniques, which limit their acceptance and actual use in highly safety-critical domains like healthcare where interpretability and reliability are of vital importance.

To counterbalance the effects of data scarcity and/or high internal variability, prior work has suggested the use of regularization to improve the model's generalization ability [9,10]. Regularization refers to techniques that reduce overfitting and constrain the complexity of a ML model. They result in a more stable learning process, smoothen the loss surface and can help to balance the bias-variance trade-off to optimize the model's performance. However, the fundamental issue of having orders of magnitudes higher number of trainable parameters in NNs compared to the number of training samples remained a bottleneck, not properly addressed in the field of DL.

Recently, a novel, parameter-efficient neural network approach has been introduced by Papp et al. [11] called the Distance-Encoding Biomorphic-Informational Neural Network (DEBI-NN) exhibiting promising results in processing small medical datasets while consolidating deep neural networks. Instead of optimizing weights, DEBI-NNs train spatial positions of their neurons consisting of soma-axon pairs, where weights are derived from the spatial distances of connected neurons. We refer to this property as *spatial plasticity*. DEBI-NNs are three-dimensional, spatial neural networks where the soma-axon pairs have both x,y,z coordinates (unknown parameters to train) in the Euclidean space. Every time a DEBI-NN is modified in terms of its neural coordinates, distances are recalculated to weights. This construct allows the network to linearly scale with the number of trainable parameters. In addition, the effect of modifying DEBI-NN coordinates is higher than modifying the same number of weights in a conventional NN, as a soma or axon modification in the Euclidean space affects all connected neurons and their calculated weights. In Ref. [11], the initial version of this architecture demonstrated performance on par with neural networks (NN), albeit, with significantly less parameters to train. Additionally, it appeared to handle class imbalance more successfully without explicit regularization. However, due to the novelty of the approach, the optimal configurations and regularizing behaviour of this network are yet to be explored. Hence, in this study, we aimed to conduct a follow-up evaluation involving a wide-range of regularization approaches to compare their effect on predictive performance in DEBI-NNs and conventional NNs.

Based on the above initial findings, we hypothesize that DEBI-NN models can achieve increased generalizability due to spatial plasticity combined with significantly fewer trainable parameters compared to their conventional fully-connected NN counterparts. We further hypothesize that DEBI-NNs have inherent regularization properties due to spatial plasticity, thus, they can further simplify model building in small medical imaging cohorts compared to conventional NNs. In order to test our hypotheses, we aimed to conduct a multi-centric comparative study involving various datasets, network layouts and regularization combinations. Therefore, we established the following objectives: (a) to collect a wide-range of cancer cohorts having variations in the type or modality of data, sample size and class imbalance ratio where data is originated from at least two sites; (b) to measure the predictive performance of multiple model configurations in both DEBI-NN and conventional NN models, considering all combinations of a set of conventional and DEBI–NN–specific regularization techniques; and (c) to assess the benefit of regularization techniques regarding predictive performance in an independent test setting in both DEBI-NN and conventional NN models.

2. Methods

For the CONSORT figure of the study, see Fig. 1. For the Checklist for Artificial Intelligence in Medical Imaging table, see Appendix E.

2.1. Datasets

All cohorts were generated by studies that were previously reviewed and approved by the appropriate local institutional research ethics committees to be involved in retrospective AI analyses. Whenever appropriate, written consents were obtained. For details regarding ethical approvals, refer to the publications of the original studies, providing the data this study utilized (see below).

2.2. Diffuse large B-cell lymphoma dataset (DLBCL)

This radiomic and clinical dataset includes data collected from 85 high-risk patients with histologically proven diffuse large B-cell lymphoma (DLBCL) with the goal of predicting two-year event-free survival. The [^{18}F]FDG PET/CT scans and relevant clinical parameters were retrospectively collected between 2014 and 2019 from two centers: 41 patients from the University of Pécs, Department of Medical Imaging (Center 1) and 44 patients from University of Kaposvár, Hungary (Center 2), which populate the train and test sets, respectively. The patient population's median age was 59 years (23–81 years) with 48.20 % ($n = 41$) of patients older than 60. The ratio of male to female patients in the cohort was 47 % ($n = 40$) and 53 % ($n = 45$), respectively. Comprehensive clinical characteristics of these patients, including treatments and biomarkers, have been described in the original study [12]. This dataset was used to evaluate predictive models for survival outcomes predicting 2-year survival (yes/no), leveraging in vivo radiomics data derived from baseline [^{18}F]FDG PET/CT and clinical parameters. For specific dataset characteristics, see Table 1.

2.3. Head and neck tumor dataset (HECKTOR)

This head and neck [^{18}F]FDG PET/CT dataset was collected and curated within the context of the 2022 MICCAI challenge (HECKTOR) [13]. The patients in the dataset were labelled as Human Papillomavirus (HPV) status positive or negative. The input data combined 8 clinical variables (age, gender, stage, treatment, etc.) and 28 image biomarker standardization initiative (IBSI) - compliant 3D radiomic features extracted from the delineated tumour volumes in both the FDG PET and CT scans. The dataset is heterogeneous, with images provided by several centers that were split into 158 training and 74 test samples.

2.4. Prostate cancer dataset (PROSTATE)

This dataset includes data from two centers including Augsburg (DE), supplying 50 [^{68}Ga]Ga-PSMA-11 (PSMA-11) PET/CT cases and 28 PSMA-11 PET/MRI cases from Vienna (AT) [14], respectively serving as train and independent test sets (see Table 1). Both sets underwent automated prostate detection (in CT and in MRI) relying on the segmentation component of the Dedicaid service (Telix Pharmaceuticals, IN, USA) hosted at MedUni Wien. This was followed by primary prostate-based IBSI radiomic feature extraction [15] from PSMA PET images relying on the IBSI-conform radiomics component of the Dedicaid service, where features having "strong" or "very strong" multi-centric consensus according to IBSI were extracted. All cases were labelled by whole-mount histopathology-provided International Society of Urological Pathology (ISUP) grades as low (ISUP ≤ 2) and high (ISUP >2) risk.

2.5. Data preprocessing

Data preprocessing was performed on the training set. The test set was solely used for performance evaluation and did not influence the modelling process. Constant features were removed before applying feature redundancy removal. Features were additionally removed if they had more than 20 % missing values. Samples were removed if labels were missing. Subsequently, z-score standardization was performed [11], followed by k-nearest neighbour feature imputation. Lastly, minimum-Redundancy Maximum-Relevance (mRMR) feature redundancy ranking and selection was done by pairwise correlation with Spearman ranking and then the feature with the lower variance was discarded for pairs with correlation coefficient >0.8 [16]. The final selected features and the z-score coefficients were also applied to the corresponding test sets.

2.6. Regularization techniques

As a baseline, conventional neural networks (NNs) were trained next to DEBI-NN models as they are good demonstrators of regularization methods utilized in traditional DL models [17].

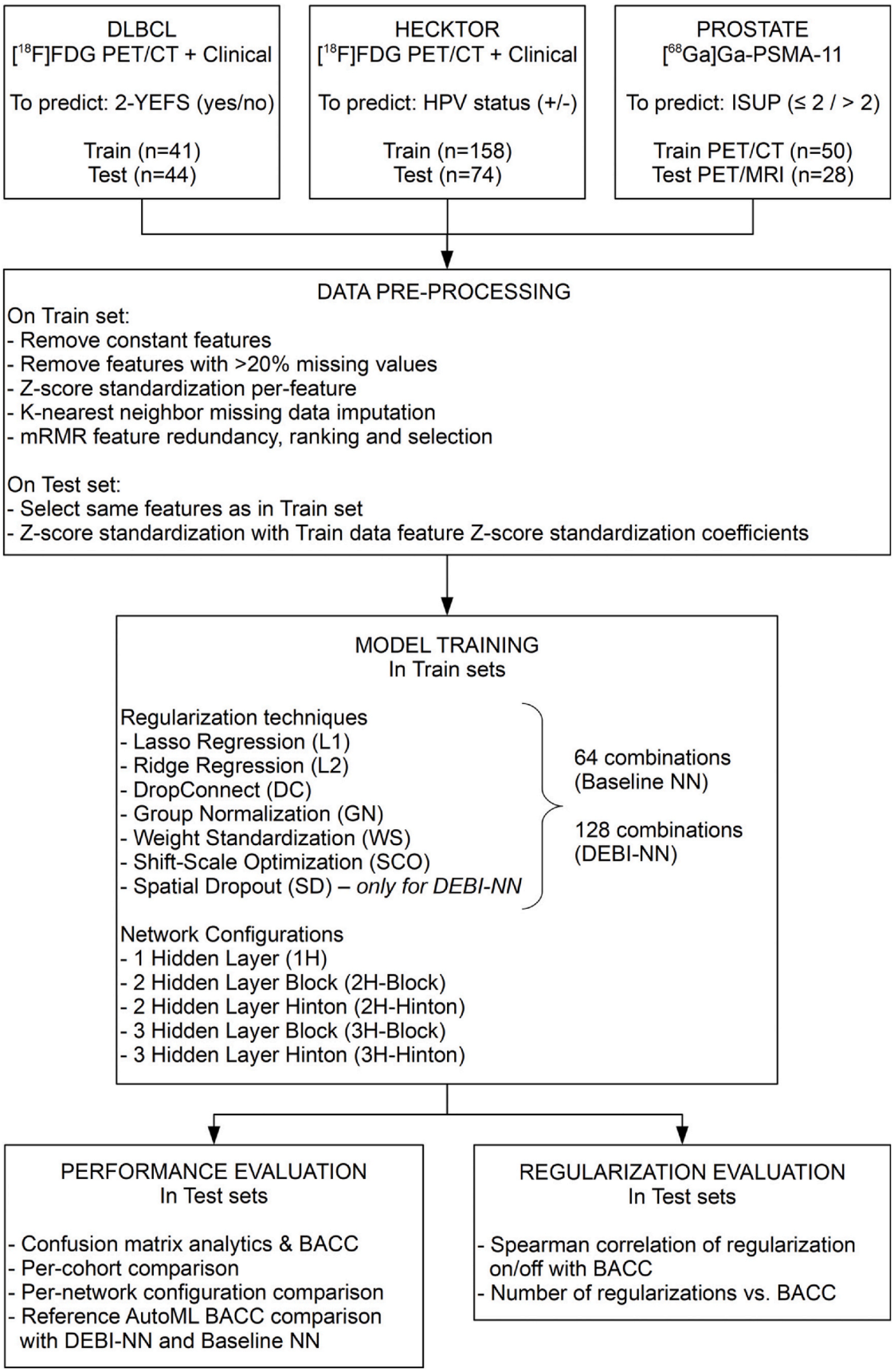
The regularization techniques for this study were selected based on the following criteria: whether they can generally be applied to both fully-connected NNs and DEBI-NNs, and their prevalence in the practice of NNs regularization. We omitted methods such as Batch Normalization [18,19], as they are not used for NNs that are not optimized with gradient descent-based methods. Consequently, they were also not utilized for DEBI-NNs. Furthermore, we utilize Spatial Dropout (SD) as a DEBI–NN–specific regularization technique [20].

The following regularization techniques were implemented in this study: L1 or Lasso regression (L1), L2 or Ridge regression (L2), Drop Connect (DC), Group Normalization (GN), Weight Standardization (WS), Shift-Scale Optimization (SCO), and the DEBI–NN–specific Spatial Dropout (SD). For further details about these techniques, see Appendix C. Note that applying both L1 and L2 together yields the so-called "Elastic Net" regularization [21], but for consistency, we regard it as a combination of methods rather than a standalone method. Some other techniques, for example WS and GN are often applied together to enhance the stability and performance of the training. For this reason, we aimed to investigate both the individual merit of the regularization techniques as well as their combined effects on the training process. Therefore, configurations were built from all possible combinations of the above methods for both NNs and DEBI-NNs, resulting in a total of $2^6 = 64$ and $2^7 = 128$ configurations (since DEBI-NNs additionally have SD), respectively. For detailed configurations, see Appendix C.

To isolate the effects of different regularization configurations and to ensure a controlled evaluation, other hyperparameters were held constant across network architectures and training procedures, as shown in Appendix B. In case of the baseline NN models, learning rate and weight decay were optimized with grid search, since these parameters significantly influence predictive performance, as detailed in Appendix B. DEBI-NNs employed no hyperparameter optimization. Furthermore, early stopping was enabled in all the experiments. Further details on the specific training and hyperparameter optimization configurations can be found in Appendix B and C.

2.7. Network configurations

The present study examined regularization behaviour across five architecture layouts, denoted as "1H", "2H-Block", "2H-Hinton", "3H-Block", and "3H-Hinton", referring to networks with 1, 2, or 3 hidden layers, respectively. "Block" and "Hinton" are different rules of hidden layer configurations as defined in Ref. [11] and detailed for each dataset in Appendix B. In short, a 1-hidden layer network has 5-times hidden neuron counts relative to the input feature counts; Block networks have the same neuron counts across hidden layers (2-times the input feature count), while Hinton networks reduce the number of neurons in consecutive hidden layers 2/3-times of the neuron count relative to the



(caption on next page)

Fig. 1. The CONSORT chart of the study. Three cancer cohorts are involved in the study, having different sample counts, imbalance ratios and clinical binary endpoints to predict. All data are dual or multi-centric, including imaging and non-imaging (clinical) features to predict from. Data pre-processing ensures that high-quality features are selected (without missing data and with high correlation with training labels to predict, while minimizing redundancy). Model training involved 6 and 7 regularization techniques for Baseline NN and for DEBI-NN, respectively, resulting in $2^6 = 64$ and $2^7 = 128$ possible configurations in total. This is combined with 5 different network configurations per regularization combination and per cohort. Performance evaluation is done in test datasets by calculating confusion matrix analytics and balanced accuracy per model. DEBI-NN, Baseline NN and AutoML models are compared. Regularization evaluation is conducted by correlating test balance accuracies and regularization techniques across cohorts and network schemes. DLBCL – Diffuse Large B-cell Lymphoma; 2-YEFS – 2-years event-free survival; HECKTOR – Head and Neck tumour cohort; HPV – Human papillomavirus; ISUP – International Society of Urological Pathology; NN – Neural Network; DEBI-NN – Distance Encoding Biomorphic NN; BACC – balanced accuracy; PET/CT – Positron Emission Tomography/Computed Tomography; MRI – Magnetic Resonance Imaging.

Table 1
Dataset characteristics. Number of samples and features refer to the properties of the datasets after Data Pre-processing. The minority class imbalance ratio describes the ratio of the minority class size relative to the total number of samples in the corresponding data. DLBCL – Diffuse Large B-Cell Lymphoma dataset; HECKTOR – Nead and Neck dataset; HVP – Human Papillomavirus; ISUP – International Society of Urological Pathology, HPV – Human papillomavirus.

	DLBCL		HECKTOR		PROSTATE	
Role	Train	Test	Train	Test	Train	Test
# Samples	41	44	158	74	50	28
# Features	17		36		31	
Minority class imbalance (%)	39 %	31.8 %	37.3 %	25.7 %	50 %	39.3 %
Classification output	2-year survival (yes/no)		HPV status (±)		ISUP grade (≤ 2/> 2)	

previous layer. Both NNs and DEBI-NNs follow these layer configurations, i.e., in a given layout, NNs and DEBI-NNs have the same number of neurons in corresponding layers.

A total of 64 and 128 different regularization configurations were tested for NNs and DEBI-NNs on 5 different network architecture layouts in such a way that all possible combinations of the regularization methods were applied on each of the five layouts, yielding a total number of 320 and 640 experiments per cohort for NNs and DEBI-NNs, respectively.

2.8. Predictive performance estimation

Predictive performance was estimated with both DEBI-NN and Baseline NN models on all three datasets, performing binary classification tasks. Predictive performance was estimated on the independent test sets of each cohort. The number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) cases was quantified. The reported metrics were sensitivity (SNS), specificity (SPC), positive predictive value (PPV), negative predictive value (NPV), accuracy (ACC), and balanced accuracy (BACC) to compare predictive performance between the baseline NN and DEBI-NN models, as well as across different configurations and layouts.

Our primary performance metric used for the analysis was balanced accuracy (BACC) as it is more accurate and comprehensive of the results for small, and especially imbalanced datasets than any one of the other metrics alone.

The models have been evaluated on the same datasets and with highly similar configurations. However, their training paradigms are fundamentally different and therefore quantitatively comparing their performance is a challenging task. To ensure fair cross-architectural comparison, we report both absolute and relative (normalized) performance differences. The absolute difference between DEBI-NN and Baseline NN balanced accuracies is denoted as Δ BACC (abs), while the mean difference is denoted as Δ Mean BACC (abs) hereafter.

Additionally, a relative performance gain metric was introduced to express the scale of the performance difference between DEBI-NN and Baseline NN calculated by normalizing gains by the baseline's performance referred to as Δ BACC (%), while the mean relative difference is referred to as Δ Mean BACC (%) hereafter.

Last, each cohort train-test configuration was used to build 100 different mixed-stacked ensemble learners (a.k.a. super-learners) automatically (AutoML) in the Dedicaid service. The 100 super-learners included Bayesian classifiers, Gaussian mixture models, random forests and support vector machines [12,22]. These 100 super-learners then provided a final prediction on the test cases by a majority vote.

2.9. Evaluation of regularization techniques

Test BACC values of each model were correlated with the presence of regularization techniques (on/off) to investigate which of these techniques individually and combined affect DEBI-NN and Baseline NN predictive performance relying on Spearman correlation. This was done per cohort and per network configuration.

Relative BACC changes per regularization and the occurrence of those BACC changes were compared between DEBI-NN and baseline NN models.

The number of regularizers present vs. test BACC was also investigated in DEBI-NNs and Baseline NNs by categorizing BACC values into groups having the same number of regularizations active. Trendlines regarding test BACC change were analysed in light of number of regularizations being on across DEBI-NN and baseline NN models.

3. Results

3.1. Predictive performance evaluation

As shown in Fig. 2, DEBI-NN yielded mean test BACC of 69.15 % (±1.44 95 %CI) on DLBCL, 68.33 % (±1.48 95 %CI) on HECKTOR, and 60.73 % (±1.82 95 %CI) on PROSTATE, respectively. In contrast, Baseline NN yielded a mean test BACC of 56.17 % (±1.67 95 %CI) on the DLBCL, 62.09 % (±1.83 95 %CI) on the HECKTOR, and 52.75 % (±1.35 95 %CI) on the PROSTATE datasets across all network architectures. For detailed results see Appendix D. For an example DEBI-NN model in 3D, see Fig. 3.

Table 2 summarizes the absolute and relative performance improvements of the DEBI-NN models over the Baseline NNs where relative improvement normalizes gains by the Baseline's performance, enabling fair cross-architecture comparisons. On the DLBCL dataset DEBI-NN demonstrates $+18.1 - 29.3$ % relative improvement over Baseline. On the HECKTOR dataset, DEBI-NN demonstrates $+4.4 - 15.7$ % relative improvement, on the PROSTATE dataset $+10.3 - 21.8$ % relative improvement. The highest absolute mean performance difference of 15.9 % occurred in the DLBCL dataset, while the lowest absolute mean performance difference of 2.6 % was observed in the HECKTOR dataset. The most significant performance gap was observed in the DLBCL dataset, with the lowest average variance, and therefore the highest

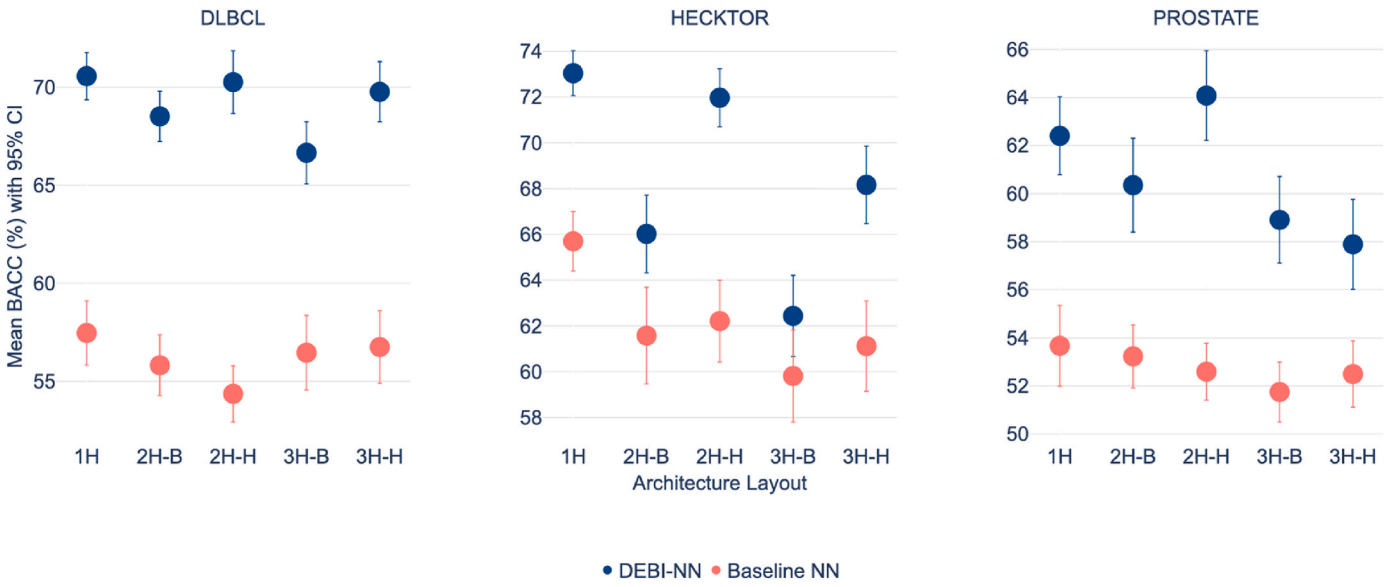


Fig. 2. Absolute Mean Balanced Accuracy (BACC (abs)) with 95 % confidence intervals (CI) for DEBI-NN and Baseline NN models across three datasets. Each subplot corresponds to a dataset (DLBCL, HECKTOR, and PROSTATE) and displays the absolute mean BACC over different configurations for the given architecture layout (1H, 2H–B (Block), 2H–H (Hinton), 3H–B, 3H–H) shown along the x-axis. For the full table displaying mean and CI values, please see Appendix D.

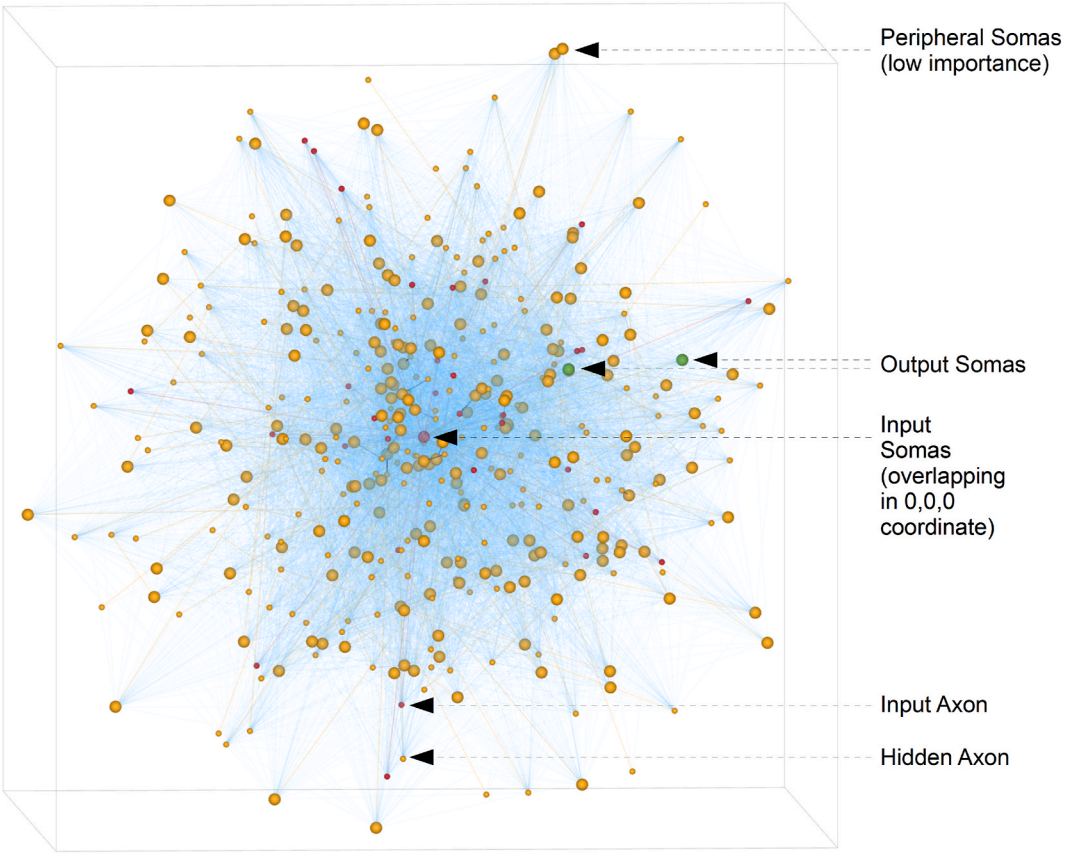


Fig. 3. Example 3D rendering of a PROSTATE DEBI-NN model. The distributions of the artificial somas and axons are centralized around the input somas (large red spheres in the middle, overlapping) due to the model being built with the so-called Singularity initialization mode. This initialization mode puts all somas and axons to the 0,0,0 Cartesian coordinate in the 3D Euclidean space, from which training iteratively modifies all trainable somas and axons until their distances (and thereby, weights) are optimal. Input somas are not trained, as their position is invariant. For the same reason, output axons are also not trained. Hidden somas and axons (yellow) as well as output somas (green) are trained. Blue lines correspond to calculated weights. When a feature in the given layer is not important, its axons tend to be positioned far from next layer somas. This phenomenon is typically observed through peripheral somas or axons. For more information about DEBI-NN parameters, see the openly-available DEBI-NN Handbook in the GitHub repository (see Access).

Table 2
Performance improvements of DEBI-NN over Baseline NN across various layouts and datasets. The table presents absolute (Δ Mean BACC (abs)) and relative (Δ Mean BACC (%)) gains in mean Balanced Accuracy (BACC). The mean BACC values were calculated for each model layout (1H, 2H-Block, 2H-Hinton, 3H-Block, 3H-Hinton) for each dataset (DLBCL, HECKTOR, PROSTATE). Positive values indicate that DEBI-NN models have higher performance compared to Baseline NNs. For the full table displaying mean values with improvement percentages, see Appendix D.

Dataset	Layout	Δ Mean BACC (abs)	Δ Mean BACC (%)
DLBCL	1H	13.1	22.8
	2H-Block	12.7	22.7
	2H-Hinton	15.9	29.3
	3H-Block	10.2	18.1
	3H-Hinton	13	22.9
HECKTOR	1H	7.3	11.2
	2H-Block	4.4	7.2
	2H-Hinton	9.8	15.7
	3H-Block	2.6	4.4
	3H-Hinton	7	11.5
PROSTATE	1H	8.7	16.3
	2H-Block	7.1	13.4
	2H-Hinton	11.5	21.8
	3H-Block	7.2	13.8
	3H-Hinton	5.4	10.3

Table 3
Comparison between DEBI-NN and Baseline NN 95 % Confidence Interval (CI) Separation and Precision Ratio. Values were calculated for each model layout (1H, 2H-Block, 2H-Hinton, 3H-Block, 3H-Hinton) for each dataset (DLBCL, HECKTOR, PROSTATE). Separation expresses the difference between the lower 95 % confidence interval (CI) bound of the DEBI-NN and the upper 95 % confidence interval bound of the Baseline NN. Positive separation values mean that the confidence intervals are non-overlapping, while negative values indicate overlapping intervals. Precision Ratio (PR) computes the relative CI widths of DEBI-NN and Baseline NN. PR values < 1.0 indicate tighter CI bounds for DEBI-NN, while values > 1.0 indicate tighter CI bounds for Baseline NNs. For the full table displaying confidence intervals, separations and precision ratios, refer to Appendix D.

Dataset	Layout	Separation	Precision Ratio
DLBCL	1H	10.3	0.7
	2H-Block	9.9	0.8
	2H-Hinton	12.9	1.1
	3H-Block	6.7	0.8
	3H-Hinton	9.6	0.8
HECKTOR	1H	5.1	0.8
	2H-Block	0.6	0.8
	2H-Hinton	6.7	0.7
	3H-Block	−1.2	0.9
	3H-Hinton	3.4	0.9
PROSTATE	1H	5.4	1
	2H-Block	3.9	1.5
	2H-Hinton	8.4	1.6
	3H-Block	4.1	1.4
	3H-Hinton	2.1	1.4

stability (as shown in Fig. 2, Tables 2 and 3). On the HECKTOR and PROSTATE datasets, the performance gain was low to moderate. Among the five network architectures, 2H-Hinton seems to be the most robust with an average absolute improvement of +12.4 % and an average relative improvement of +22.3 % across all datasets. 1H models also performed particularly well on all datasets, while 3H-Hinton models were successful on the DLBCL and HECKTOR datasets.

Table 3 summarizes the separation and precision ratio (PR) corresponding to the 95 % Confidence Intervals (CIs). Regarding overall performance consistency, DEBI-NN is most decisively superior on the DLBCL dataset with a CI separation of ≥ 6.7 % and an overall clear advantage on the HECKTOR and PROSTATE datasets (as shown in Table 2) but with varying effect sizes. All but one architecture layouts exhibit non-overlapping confidence intervals with a separation > 0.0 (except for HECKTOR 3H-Block models). Most precision ratios in the DLBCL and HECKTOR datasets are < 1.0 (except DLBCL 2H-Hinton), indicating tighter confidence intervals for the DEBI-NN models. On the PROSTATE dataset, the Baseline NN bounds are tighter.

The highest-performing DEBI-NN models with test BACC were 2H-Hinton with 84.5 % in DLBCL, 1H and 3H-Hinton with 80 % in HECKTOR and 3H-Block with 80.5 % in PROSTATE.

In contrast, the highest-performing Baseline NN models were 3H-Block with 71.9 % in DLBCL, 2H-Block with 77.3 % in HECKTOR and 1H with 77.3 % in PROSTATE (Fig. 4).

The reference AutoML evaluations yielded a test BACC of 79 % for DLBCL, 75.5 % for HECKTOR and 54.5 % for PROSTATE.

Fig. 2 demonstrates that the highest-performing DEBI-NN model instances consistently outperform the highest-performing Baseline NNs across all datasets and architecture layouts by an average of 12.0 %, 3.8 %, and 5.5 % improvement for the DLBCL, HECKTOR, and PROSTATE datasets, respectively. The highest difference of 15.9 % between the best-performing models was exhibited by 2H-Hinton on the DLBCL dataset, while the smallest difference of 0.2 % was observed between the 1H models on the PROSTATE dataset.

3.2. *Evaluation of regularization techniques*

Fig. 5 reveals how different regularization methods impact the Baseline NN and DEBI-NN absolute BACC performance measured by the Spearman correlation coefficient for each layout and regularization technique across all datasets. In general, most correlation values tend to be neutral or weakly correlated with performance. The highest positive values show a weak to moderate correlation, while the highest negative values show a moderate to strong correlation.

In case of DEBI-NN, there are 9/35, 14/35, and 15/35 positively correlated cells for the DLBCL, HECKTOR, and PROSTATE datasets respectively, yielding a total of 38/105 positively correlated values. In case of Baseline NN, there are 3/30, 13/30, 15/30 positively correlated cells, yielding a total of 31/90 positively correlated values. See Fig. 6 for the distribution of correlations of predictive performances with individual regularization techniques. Fig. 7 represents the predictive performance differences between best regularization vs. no regularization models in DEBI-NN and baseline NN models.

Fig. 8 reveals that DEBI-NNs benefit less and less, while baseline NNs benefit more and more form the presence of more than one regularization techniques. However, on average, DEBI-NN models have a higher predictive performance than Baseline NN models have even if the latter one utilizes a high number of regularizations. In particular, the highest average test BACC across all models in DEBI-NNs was 61.67 % (±1.557 95 %CI) having 1 regularizations active, while the test BACC tendency was systematically decreasing down to 56.38 % (±3.71 95 %CI) with 7 regularizations being active. In contrast, Baseline NNs has a consistent test BACC increase starting from 53.06 % (±3.0 95 %CI) with no regularization up to 59.93 % (±3.71 95 %CI) with 6 regularizations. For details see Appendix D.

4. Discussion

In this study, we conducted a comprehensive investigation into the effects of various regularization techniques on the Distance-Encoding Biomorphic-Informational Neural Network (DEBI-NN) architecture. We compared its performance against Baseline NNs across three distinct, multi-centric small medical imaging datasets, utilizing several network layouts and combinations of regularization configurations. Our primary hypothesis was that DEBI-NN models, owing to their inherent spatial plasticity and significantly fewer trainable parameters, would exhibit enhanced generalizability and potentially a reduced need for explicit regularization compared to conventional NNs.

The results largely support our hypothesis. A key finding is the consistent superior predictive performance of DEBI-NN models over

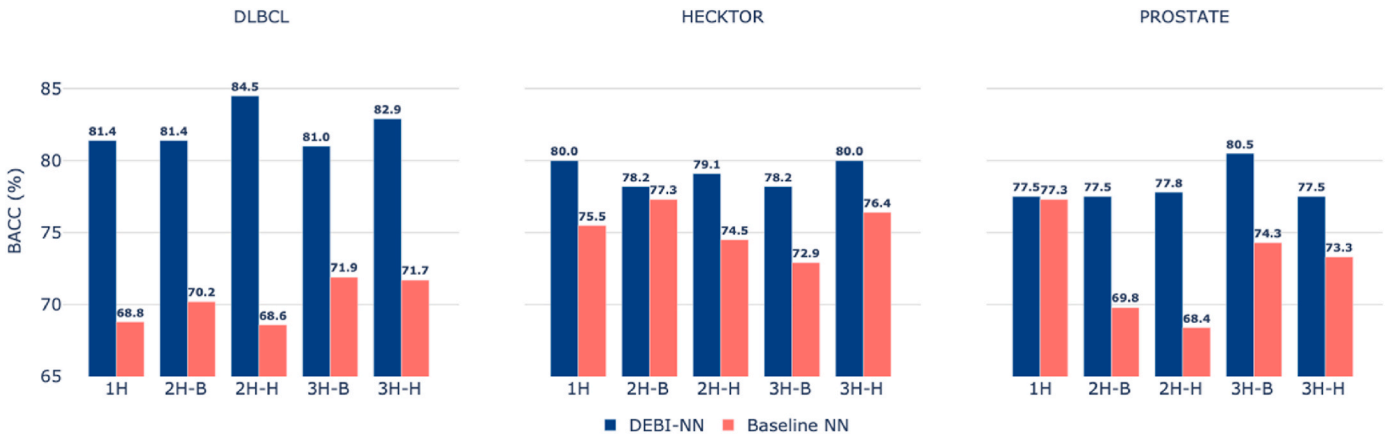


Fig. 4. Comparison of best performing DEBI-NN and Baseline NN models across five architecture layouts and three datasets measured by absolute balanced accuracy (BACC). Each column corresponds to the highest performing model instance for the given layout (1H, 2H-Block, 2H-Hinton, 3H-Block, 3H-Hinton) and dataset: DLBCL, HECKTOR, PROSTATE. The BACC values are presented as percentages. For full tables containing the five best and worst performing DEBI-NN and Baseline NN models for DLBCL, HECKTOR, and PROSTATE datasets, refer to Appendix D.

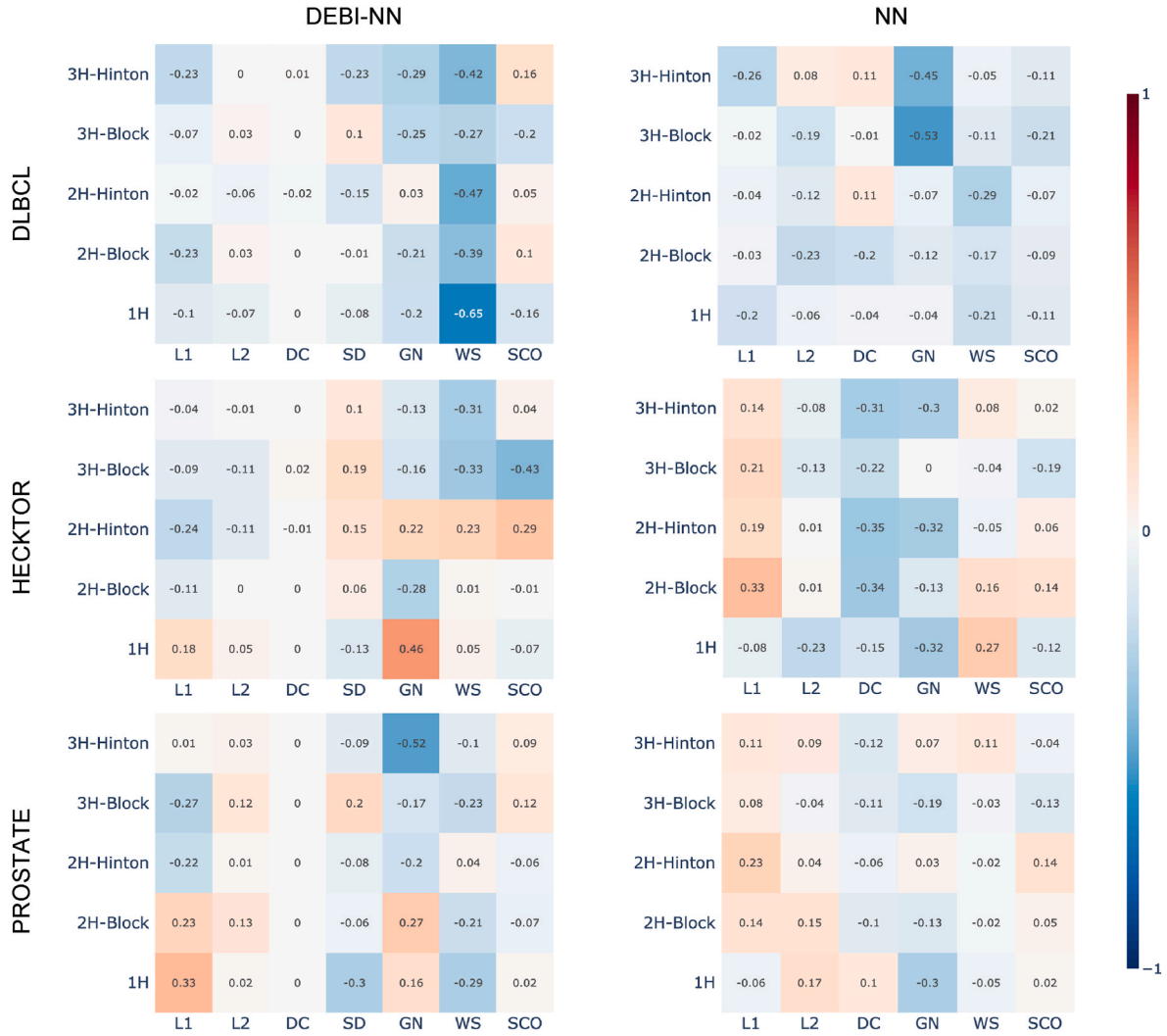


Fig. 5. Spearman correlation coefficients between various regularization techniques active and the balanced accuracy (BACC) metric for DEBI-NN and Baseline NN models across three datasets showing the general effect of regularization techniques on performance. Each heatmap row represents a distinct model architecture layout (1H, 2H-Block, 2H-Hinton, 3H-Block, 3H-Hinton), while each column corresponds to a specific regularization technique. Red tones are indicating positive (regularization increases predictive performance), while blue tones are indicating negative correlations (regularization decreases predictive performance). L1 – Lasso regression; L2 – Ridge regression; DC – Drop Connect; GN – Group Norm; WS – Weight Standardization; SCO – Shift-Scale Optimization; SD – Spatial Dropout (DEBI-NN specific).

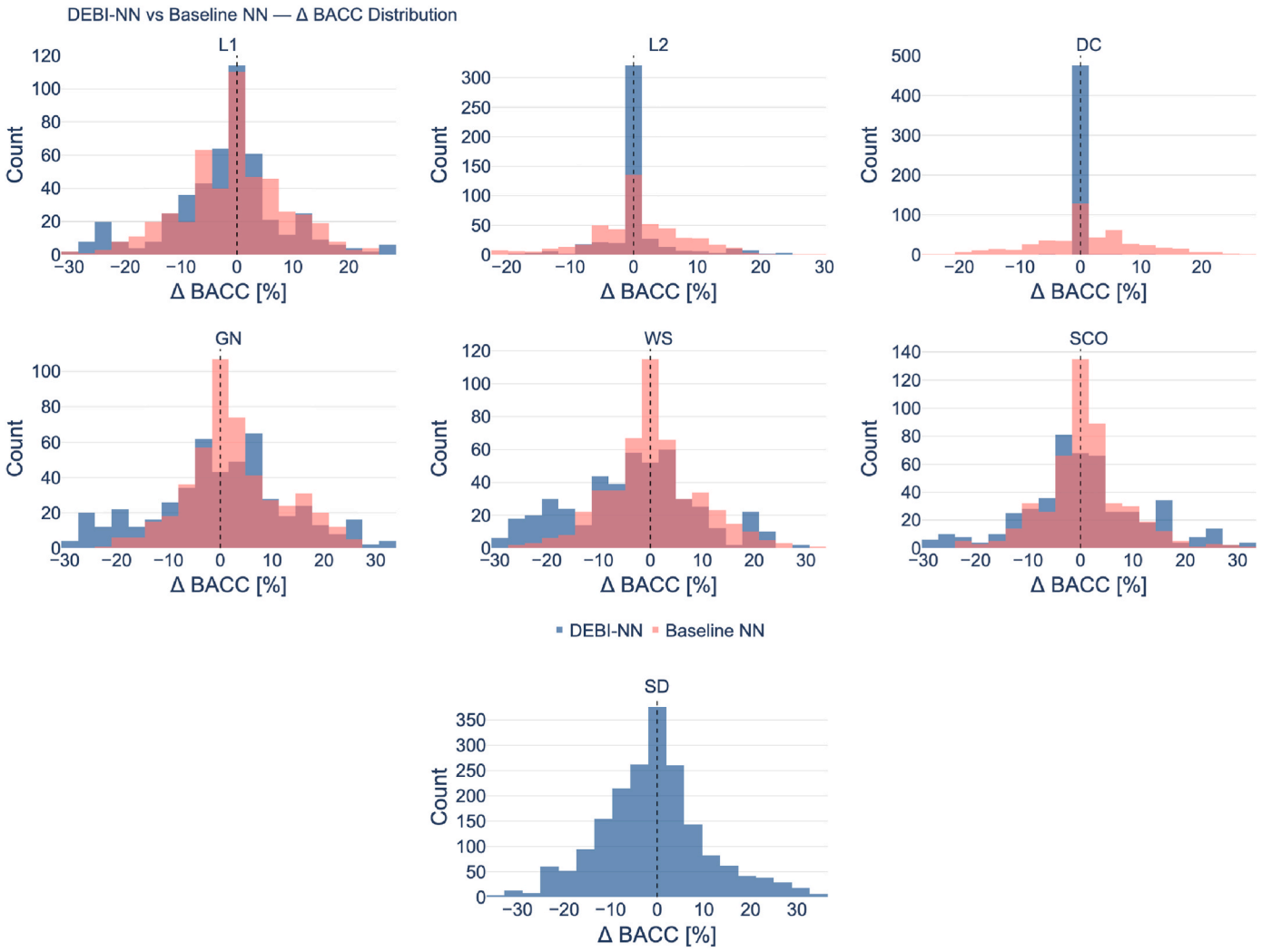


Fig. 6. The histogram of test predictive performance changes in light of the given regularization technique being activated overlaid between DEBI-NN and Baseline NN models. The vertical dashed line (value zero on the x-axis) denotes no effect of the regularization technique. X-axes denote the delta change of BACC when the given regularization is active. Negative X-axis values imply that the regularization technique decreases test BACC, and positive values on the x-axis represent a positive effect of the given regularization on predictive performance. Y-axes values represent the occurrence of the given test balance accuracy delta across all cohorts and model instances. L2 and DC appear to have minimal to no effect on DEBI-NN predictive performances, while GN, WS and SCO represent noticeable effects on predictive performance. Looking at the x-axes of GN, WS and SCO reveals that DEBI-NNs on average do not benefit from the presence of these regularization techniques as much as Baseline NNs do. Spatial Dropout (CD), a DEBI-NN specific regularization technique represents a high test BACC Spread in light of having SD activated, implying that its effect is not generic, but model-specific which is in line with the findings of [20]. L1 – Lasso regression; L2 – Ridge regression; DC – Drop Connect; GN – Group Norm; WS – Weight Standardization; SCO – Shift-Scale Optimization; SD – Spatial Dropout (DEBI-NN specific).

Baseline NNs. As demonstrated in Fig. 2 and Table 2, DEBI-NNs achieved higher mean Balanced Accuracy (BACC) across all tested network architectures and datasets. Specifically, DEBI-NNs showed an average absolute BACC improvement of 9.1 % and relative BACC improvement of 16.1 % over Baseline NNs, with the most substantial gains observed in the DLBCL dataset (Fig. 2). This dataset also showed the highest stability and CI separation for DEBI-NN performance (Table 3). The DLBCL dataset's characteristics – being the smallest in terms of training samples, having the fewest number of features, and exhibiting significant class imbalance (especially in the test set with 31.8 %) – create a particularly challenging task for conventional DL models, and an opportunity to demonstrate the key strengths of DEBI-NNs, such as efficiency with small data, reduced parameter space, and robustness to imbalance. At the same time, we observed minor variability in average performance across datasets (Appendix D). We believe these discrepancies primarily reflect differences in the underlying classification tasks, including the degree of class imbalance, label noise, and overlapping feature distributions, all of which inherently affect discriminability. Importantly, DEBI-NNs consistently outperformed Baseline NNs, even if the magnitude of the improvement varied depending on dataset-specific properties.

Furthermore, the best-performing individual DEBI-NN model instances consistently surpassed their Baseline NN counterparts across all configurations (Fig. 4) without explicit hyperparameter tuning. This suggests that DEBI-NNs demonstrate significantly lower sensitivity to hyperparameter configurations compared to Baseline NNs where hyperparameter search was performed on learning rate and weight decay. This observation is consistent with machine learning literature stating that gradient descent-based methods are generally more sensitive to hyperparameter tuning than genetic algorithm-based ones [23–25]. Furthermore, the highest-performing DEBI-NN models also systematically outperformed the AutoML super-learner models, demonstrating that DEBI-NNs can be viable DL alternatives of such methods in future studies, operating in small, imbalanced medical data. Given the generic trends within the field of AI that aim to increase model complexity, currently, there is no known similar work which intends to

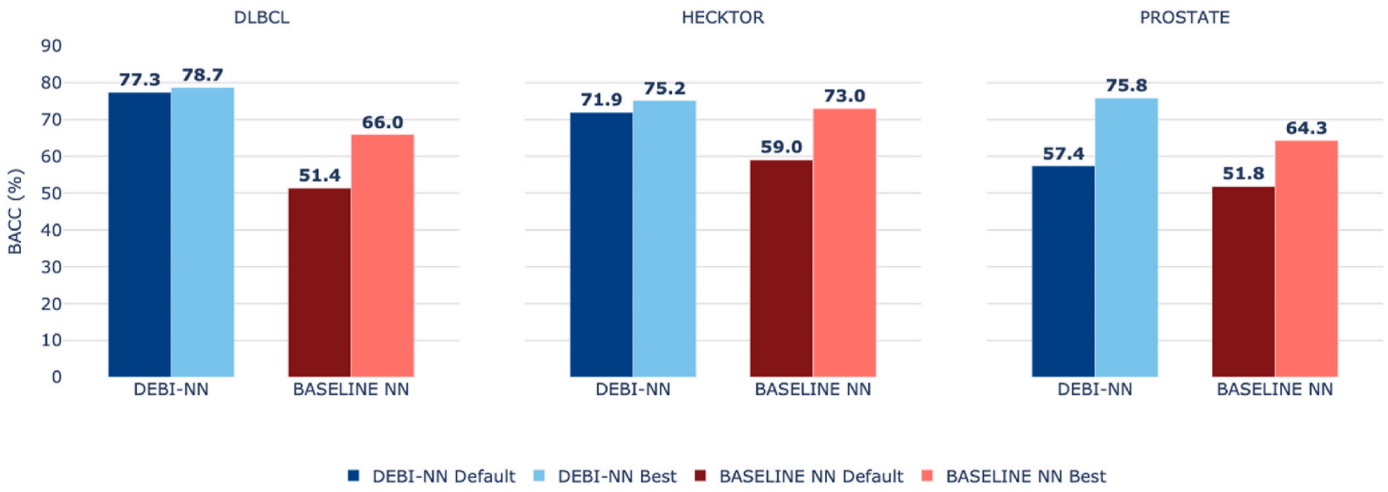


Fig. 7. Comparison of no regularization (Default) and best regularization (Best) configurations in DEBI-NN and Baseline NN models based on average test balanced accuracy (BACC) performance across layouts for the given configuration.

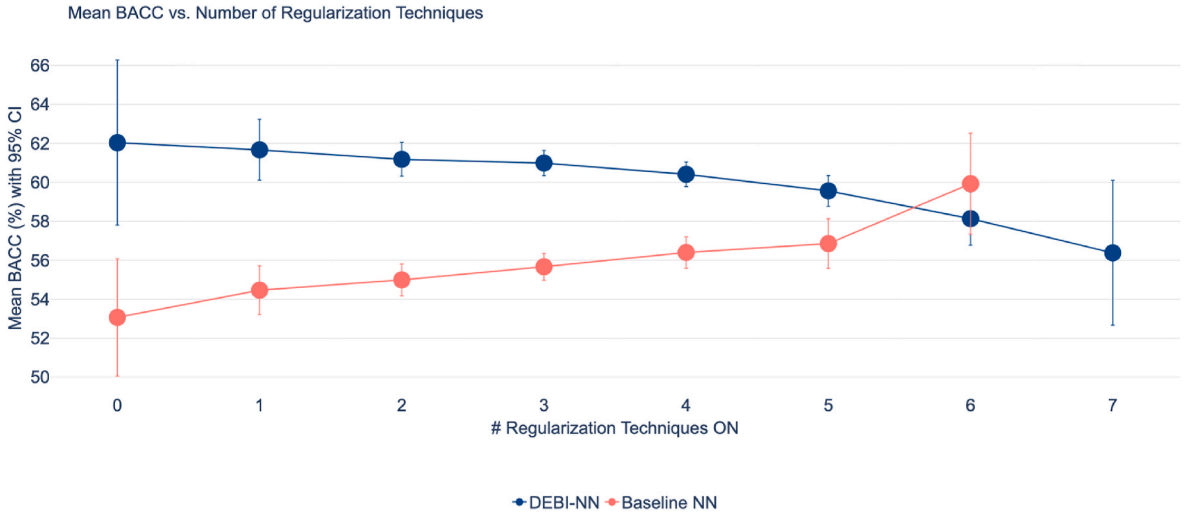


Fig. 8. Mean Balanced Accuracy (BACC) with 95 % CI vs. Number of Regularization Techniques Turned ON. The numbers 0–7 represent the number of regularization techniques applied in a group of configurations. Note that for DEBI-NN, there are 7 regularization options, for Baseline NNs, there are only 6. The mean BACC with 95 % CI was calculated over each group of regularization technique. Blue markers represent DEBI-NN, while red markers are for Baseline NN models.

significantly reduce neural network parameter counts and overall training complexity while increasing predictive performance at the same time. On that note, prior works that succeeded to utilize DL in hybrid imaging data mostly focus on segmentation tasks [1,26–28], but typically fail to achieve high performance with classification tasks [1,2,29]. The reason is that while segmentation can cut up the input training image to small parts for identifying whether a given voxel and its direct surrounding belongs to suspicious lesions - thereby dramatically increasing the number of samples to train on -, predicting a patient-level information (e.g. outcome) requires to analyse either the whole body or a complex organ associated to the prediction of the given clinical outcome. This reduces the number of training samples to the number of patients, which in return makes the utilization of DL challenging.

Our most outstanding findings indicate that DEBI-NNs can change this trend, as they possess intrinsic self-regularizing properties. As highlighted in the analysis of regularization techniques (Fig. 8), DEBI-NN models often achieve optimal or near-optimal performance with minimal or no explicit regularization. Fig. 8 reveals that the gap between the "Default" DEBI-NN configuration (with no regularization) and the best-performing configuration is generally smaller than in case of Baseline NNs. This however does not mean that the absolute best-performing configuration is the no-regularization (Default) configuration. However, it is generally observed that the worst-performing DEBI-NN models had multiple regularizers active, while the best performing DEBI-NN configurations were less regularized.

Delving deeper into specific regularization techniques, L1 regularization was a notable exception, appearing in the best-performing DEBI-NN configurations across different datasets and layouts (Fig. 6). This suggests that the feature-selecting property of L1 norm might complement the DEBI-NN architecture. Consequently, many other standard regularization techniques showed neutral, weak, or even negative correlations with DEBI-NN performance (

Fig. 5). For example, Weight Standardization (WS) exhibited a strong negative correlation (−0.65) with performance for 1H DEBI-NN models on the DLBCL dataset. The DEBI–NN–specific Spatial Dropout (SD) did not consistently emerge as a top-performing regularizer on its own, implying that the spatial plasticity of DEBI-NNs might already provide similar benefits to what SD aims to achieve. For Baseline NNs, the pattern of regularization impact was also mixed, though generally, they are expected to benefit more from regularization. However, on the DLBCL dataset, 90 % of the cases represented as cells of the top right heatmap on Fig. 5 were negatively correlated with Baseline NN

performance, potentially indicating sensitivities to specific regularizer settings or interplay between different regularizers when applied to small datasets. This might serve as an explanation for the presence of many neutrally to negatively correlated cells in the heatmaps, thus it is important to observe individual regularization configurations to determine which ones result in robust performance.

The architectural design of DEBI-NNs, characterized by training spatial positions of neurons instead of direct weights, leading to spatial plasticity and linear scaling with trainable parameters, might explain the above observations. This parameter efficiency is a significant advantage when fitting to small datasets, as it reduces the model's capacity to overfit, thereby diminishing the reliance on external regularization methods. The study by Papp et al. introducing DEBI-NNs [11] demonstrated that DEBI-NNs are capable of modelling complex, non-linear relationships in the data, thereby exhibiting fundamental properties of NNs. Our work significantly expands on these initial findings by: (a) performing a multi-centric comparative study on diverse cancer cohorts; (b) systematically evaluating an extensive suite of regularization techniques and their combinations; and (c) conducting independent validation where training and testing data originated from distinct centers.

Despite these promising findings, our study has several limitations. First of all, our study utilized and compared DEBI-NNs only against fully-connected NNs and avoided the use of current state-of-the-art DL as well as conventional ML models as baselines. However, since we used radiomic datasets, the choice of the baseline networks was justified and in line with the objectives of the study. The size and imbalanced nature of the datasets present major limitations for larger overparameterized DL models requiring the use of techniques such as data augmentation, transfer learning, pre-training, etc. which is out of the scope of the present study. The use of simple fully-connected layers as part of both DEBI-NNs and the baseline networks allowed us to establish a simple setting for fair comparison of different regularization techniques. This design choice was made to avoid further interplay with more advanced architectural solutions and to ensure that we can implement, test, and evaluate a diverse set of regularization techniques in a standardized manner on the same architectures where their effects can be best observed and compared.

Second, we also note that the chosen datasets do not meet the empirical heuristic of 10–15 events per feature [30]. Such constraints are common in radiomic and outcome prediction studies in oncology and represent one of the main reasons why deploying AI in these contexts is so challenging. This imbalance likely contributes to the variability in NN performance across architectures and datasets, by increasing the risk of overfitting and amplifying sensitivity to dataset-specific characteristics. Importantly, this limitation is not unique to DEBI-NNs but affects nearly all ML methods applied in small-sample, high-dimensional biomedical settings, as consistently reported in the literature. What our results demonstrate is that, even under these unfavourable conditions, DEBI-NNs consistently outperform Baseline NNs. We attribute this robustness to their parameter-efficient design and inherent spatial regularization, which mitigate - though cannot fully eliminate - the risks imposed by low patient-to-feature ratios. We therefore view these findings as highlighting both a limitation and an opportunity: while small sample sizes do restrict generalizability, the ability of DEBI-NNs to maintain superior performance under exactly these conditions suggests they may be particularly well-suited for data-limited, imbalanced clinical imaging scenarios. It is important to note, however, that the present findings should be interpreted as a methodological proof-of-concept under data-constrained conditions rather than as an immediately deployable clinical tool.

Third, we acknowledge that our investigation was not complete in testing and comparing all existing regularization techniques used in present-day NNs, since there exists a vast number of such techniques, many used specifically only in certain special architectures such as convolutional, spatio-temporal, graph NNs, etc. We selected a diverse set of regularization techniques that have been most popular and most successful in DL literature to date, all of which can be implemented in fully-connected layers, ensuring a fair comparison between DEBI-NN and the baselines. In addition to testing already existing regularization techniques, the present study analysed the effect of Spatial Dropout specifically developed for the DEBI-NN architecture [20]. Moreover, the study design, focusing on combinations of regularizers, also makes it challenging to fully isolate the precise impact of each individual regularizer within a combined setting, however, this allowed us to test more realistic scenarios where often multiple techniques are used together, and allowed us to investigate their combined effects.

Furthermore, while in case of Baseline NNs, learning rate and weight decay were tuned considering the high sensitivity of these hyperparameters in gradient descent-based optimizers, the hyperparameters of DEBI-NN models were not tuned. Even without hyperparameter optimization though, DEBI-NN models yielded strong performance with the chosen configuration, overperforming the Baseline models. A full hyperparameter optimization of DEBI-NNs might unveil further performance enhancements or even slightly different regularization patterns. However, the single high-performing hyperparameter configuration for DEBI-NNs proved sufficient on the chosen datasets in the present study. Furthermore, tuning the hyperparameters of specific regularization techniques could also enhance their efficacy. When choosing hyperparameter settings for regularization techniques or network schemes, we followed published guidelines on NN regularization and historically successful values observed in our prior DEBI-NN studies.

Finally, we acknowledge that BACC, while appropriate for methodological comparison, does not fully capture calibration, decision-analytic performance, or clinical utility. Future work aiming toward clinical translation will require adopting a broader evaluation framework. The focus of the present study, however, was on systematically comparing the effect of regularization strategies across datasets and architectures in a controlled manner, where BACC was the most suitable and interpretable choice which explicitly accounts for class imbalance by averaging sensitivity and specificity.

Our vision for future work is twofold [1]: to extend the DEBI scheme to architectures beyond fully connected layers, such as CNNs or Vision Transformers, which are widely used in foundation model development, and [2] to explore the integration of DEBI-NNs within multi-modal, multi-task learning frameworks to increase clinical relevance. We believe this trajectory will allow DEBI-NNs to serve not only as efficient stand-alone classifiers in constrained scenarios, but also as building blocks in more general-purpose, clinically applicable deep learning systems. In addition, the theoretical foundations of DEBI-NN's spatial plasticity and its direct link to generalization and regularization effects should be further researched.

According to the above findings, the demonstrated systematic robustness and the reduced need for a combination of regularization strategies point towards the possibility that DEBI-NNs may be ideal candidates to enable the usage of DL in small, imbalanced medical imaging datasets. This, together with the already demonstrated abilities to increase model interpretability in DEBI-NN models due to their spatial nature [20], represents the potential to make DEBI-NNs ideal candidates for both building and explaining NN models in clinical use-case scenarios within the field of medical imaging, including Nuclear Medicine. For accessing the DEBI-NN solution including an extensive handbook and example datasets, see our repository under Access.

5. Conclusions

This study provides strong evidence that DEBI-NN is a highly effective approach for predictive modelling on small, challenging medical imaging datasets. They consistently outperformed traditional neural networks and, critically, demonstrated a significantly reduced dependency on explicit regularization techniques, often achieving peak performance with minimal or no added regularizers.

6. Access

For accessing our work, see our repository: https://github.com/lpapp-muw/DEBI-NN.

This repository contains the DEBI-NN solution, all execution data with their accompanying configuration files, as well as all the raw execution results and log files.

In addition, we prepared a 34-page Handbook titled "Mastering Distance-Encoding Biomorphic-Informational Neural Networks - The DEBI-NN Handbook", which is written support the community to train, evaluate as well as to observe own DEBI-NN networks within their own research.

The handbook is accessible on Zenodo: https://zenodo.org/records/15828851.

CRediT authorship contribution statement

Boglarka Ecsedi: Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Conceptualization. **Amine Boukhari:** Writing – review & editing, Methodology, Data curation, Conceptualization. **Clemens P. Spielvogel:** Writing – review & editing, Methodology, Formal analysis, Data curation. **David Haberl:** Writing – review & editing, Software, Methodology, Formal analysis, Data curation. **Zsombor Ritter:** Writing – review & editing, Resources, Data curation. **Ralph A. Bundschuh:** Writing – review & editing, Resources, Data curation. **Constantin Lapa:** Writing – review & editing, Resources, Data curation. **Marcus Hacker:** Writing – review & editing, Resources, Formal analysis, Data curation. **Mathieu Hatt:** Writing – review & editing, Resources, Methodology, Investigation, Conceptualization. **Laszlo Papp:** Writing – review & editing, Validation, Supervision, Software, Resources, Project administration, Methodology, Conceptualization.

Declaration of competing interest

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

Laszlo Papp reports a relationship with Telix Pharmaceuticals Limited that includes: consulting or advisory. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eanmi.2025.100008.

Data availability

Data and executables are openly-accessible in the public repository of the study: https://github.com/lpapp-muw/DEBI-NN

References

[1] Hatt M, Krizsan AK, Rahmim A, Bradshaw TJ, Costa PF, Forgacs A, et al. Joint EANM/SNMMI guideline on radiomics in nuclear medicine. Eur J Nucl Med Mol Imaging 2023 Jan 3;50(2):352–75 [Internet], https://link.springer.com/10.1007/s00259-022-06001-6.

[2] Piffer S, Ubaldi L, Tangaro S, Retico A, Talamonti C. Tackling the small data problem in medical image classification with artificial intelligence: a systematic review. Prog Biomed Eng [Internet] 2024 Jul 1;6(3):032001. Available from: https://iopscience.iop.org/article/10.1088/2516-1091/ad525b.

[3] Haberl D, Spielvogel CP, Jiang Z, Orlhac F, Iommi D, Carrió I, et al. Multicenter PET image harmonization using generative adversarial networks. Eur J Nucl Med Mol Imaging [Internet] 2024 Jul 2;51(9):2532–46. Available from: https://link.springer.com/10.1007/s00259-024-06708-8.

[4] Mali SA, Ibrahim A, Woodruff HC, Andrearczyk V, Müller H, Primakov S, et al. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. J Pers Med [Internet] 2021 Aug 27;11(9):842. Available from: https://www.mdpi.com/2075-4426/11/9/842.

[5] Kshatri SS, Singh D. Convolutional neural network in medical image analysis: a review. Arch Comput Methods Eng [Internet] 2023 May 1;30(4):2793–810. Available from: https://link.springer.com/10.1007/s11831-023-09898-w.

[6] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. Available from: http://arxiv.org/abs/2010.11929; 2021 Jun 3.

[7] Ma D, Hosseinzadeh Taher MR, Pang J, Islam NU, Haghighi F, Gotway MB, et al. In: Benchmarking and boosting transformers for medical image classification; 2022. p. 12–22. Available from: https://link.springer.com/10.1007/978-3-031-16852-9_2.

[8] Hosseinzadeh Taher MR, Haghighi F, Feng R, Gotway MB, Liang J. In: A systematic benchmarking analysis of transfer learning for medical image analysis; 2021. p. 3–13. Available from: https://link.springer.com/10.1007/978-3-030-87722-4_1.

[9] Borisov V, Leemann T, Seßler K, Haug J, Pawelczyk M, Kasneci G. Deep neural networks and tabular data: a survey. Available from: http://arxiv.org/abs/2110.01889; 2021 Oct 5.

[10] Wan L, Zeiler M, Zhang S, Le Cun Y, Fergus R. Regularization of neural networks using DropConnect. In: Dasgupta S, McAllester D, editors. Proceedings of the 30th international conference on machine learning, vol. 28. Atlanta, Georgia, USA: PMLR: Proceedings of Machine Learning Research; 2013. p. 1058–66 [Internet], https://proceedings.mlr.press/v28/wan13.html.

[11] Papp L, Haberl D, Ecsedi B, Spielvogel CP, Krajnc D, Grahovac M, et al. DEBI-NN: Distance-encoding biomorphic-informational neural networks for minimizing the number of trainable parameters. Neural Networks [Internet] 2023 Oct;167:517–32. Available from: https://linkinghub.elsevier.com/retrieve/pii/S089360802300446X.

[12] Ritter Z, Papp L, Zámbó K, Tóth Z, Dezső D, Veres DS, et al. Two-year event-free survival prediction in DLBCL patients based on in vivo radiomics and clinical parameters. Front Oncol 2022 Jun 8;12 [Internet], https://www.frontiersin.org/articles/10.3389/fonc.2022.820136/full.

[13] Andrearczyk V, Oreiller V, Abobakr M, Akhavanallaf A, Balermpas P, Boughdad S, et al. In: Overview of the HECKTOR challenge at MICCAI 2022: automatic head and neck tumor segmentation and outcome prediction in PET/CT; 2023. p. 1–30. Available from: https://link.springer.com/10.1007/978-3-031-27420-6_1.

[14] Papp L, Spielvogel CP, Grubmüller B, Grahovac M, Krajnc D, Ecsedi B, et al. Supervised machine learning enables non-invasive lesion characterization in primary prostate cancer with [68Ga]Ga-PSMA-11 PET/MRI. Eur J Nucl Med Mol Imaging 2020 Dec 19;48:1795–805 [Internet], http://link.springer.com/10.1007/s00259-020-05140-y.

[15] Zwanenburg A, Leger S, Vallières M, Löck S, Initiative for the IBS. Image biomarker standardisation initiative. arXiv [Internet]. 2016;(November). Available from: http://arxiv.org/abs/1612.07003.

[16] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, et al. Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access 2016;4(October):7940–57.

[17] Kadra A, Lindauer M, Hutter F, Grabocka J. Well-tuned simple nets excel on tabular datasets. Available from: http://arxiv.org/abs/2106.11189; 2021 Nov 5.

[18] Santurkar S, Tsipras D, Ilyas A, Madry A. How does Batch normalization help optimization?. Available from: http://arxiv.org/abs/1805.11604; 2018 May 29.

[19] Wu Y, He K. Group normalization. Available from: http://arxiv.org/abs/1803.08494; 2018 Mar 22.

[20] Boukhari A, Ecsedi B, Abdallah N, Haberl D, Spielvogel C, Papp L, et al. Three-dimensional visualization of DEBI-NNs for supporting interpretability of predictive models relying on neural networks. In: XX-th international conference on the use of computers in radiation therapy (ICCR); 2024 [Internet], https://www.iccr2024.org/papers/526089.pdf.

[21] Wu Y, Liu B, Wu W, Lin Y, Yang C, Wang M. Grading glioma by radiomics with feature selection based on mutual information. J Ambient Intell Humaniz Comput [Internet] 2018;9(5):1671–82. https://doi.org/10.1007/s12652-018-0883-3.

[22] Hasimbegovic E, Papp L, Grahovac M, Krajnc D, Poschner T, Hasan W, et al. A sneak-peek into the physician's brain: a retrospective machine learning-driven investigation of decision-making in tavr versus savr for young high-risk patients with severe symptomatic aortic stenosis. J Pers Med 2021 Oct 22;11(11):1062 [Internet], https://www.mdpi.com/2075-4426/11/11/1062.

[23] Choi D, Shallue CJ, Nado Z, Lee J, Maddison CJ, Dahl GE. On empirical comparisons of optimizers for deep learning. Available from: http://arxiv.org/abs/1910.05446; 2020 Jun 16.

[24] Ojha V, Timmis J, Nicosia G. Assessing ranking and effectiveness of evolutionary algorithm hyperparameters using global sensitivity analysis methodologies. Swarm Evol Comput [Internet] 2022 Oct;74:101130. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2210650222001006.

[25] Morse G, Stanley KO. Simple evolutionary optimization can rival stochastic gradient descent in neural networks. In: Proceedings of the genetic and evolutionary computation conference 2016 [internet]. New York, NY, USA: ACM; 2016. p. 477–84. Available from: https://dl.acm.org/doi/10.1145/2908812.2908916.

[26] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021 Feb 7;18(2):203–11 [Internet], http://www.nature.com/articles/s41592-020-01008-z.

[27] Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal 2019 Sep;1(6):e271–97.

[28] Capobianco N, Meignan MA, Cottereau A-S, Vercellino L, Sibille L, Spottiswoode B, et al. Deep learning FDG uptake classification enables total metabolic tumor volume estimation in diffuse large B-cell lymphoma. J Nucl Med 2020 Jun 12;120: 242412 [Internet], http://jnm.snmjournals.org/lookup/doi/10.2967/jnumed.120.242412.

[29] Mulliqi N, Blilie A, Ji X, Szolnoky K, Olsson H, Boman SE, et al. Foundation models – A panacea for artificial intelligence in Pathology?. Available from: http://arxiv.org/abs/2502.21264; 2025 Feb 28.

[30] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning [internet]. New York, NY: Springer New York; 2009 (Springer Series in Statistics). Available from: http://link.springer.com/10.1007/978-0-387-84858-7.