Contrastive Masked Transformers for Forecasting Renal Transplant Function

Leo Milecki, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, Maria Vakalopoulou

Contrastive Masked Transformers for Forecasting Renal Transplant Function

Leo Milecki[1](✉), Vicky Kalogeiton[2], Sylvain Bodard[3,6], Dany Anglicheau[4,6], Jean-Michel Correas[3,6], Marc-Olivier Timsit[5,6], and Maria Vakalopoulou[1]

[1] MICS, CentraleSupelec, Paris-Saclay University, Inria Saclay, France
leo.milecki@centralesupelec.fr
[2] LIX, École Polytechnique/CNRS, Institut Polytechnique de Paris, France
[3] Department of Adult Radiology, Necker Hospital, LIB, France
[4] Department of Nephrology and Kidney Transplantation, Necker Hospital, France
[5] Department of Urology, HEGP, Necker Hospital, France
[6] UFR Médecine, Paris-Cité University, France

Abstract. Renal transplantation appears as the most effective solution for end-stage renal disease. However, it may lead to renal allograft rejection or dysfunction within $15\% - 27\%$ of patients in the first 5 years post-transplantation. Resulting from a simple blood test, serum creatinine is the primary clinical indicator of kidney function by calculating the Glomerular Filtration Rate. These characteristics motivate the challenging task of predicting serum creatinine early post-transplantation while investigating and exploring its correlation with imaging data. In this paper, we propose a sequential architecture based on transformer encoders to predict the renal function 2-years post-transplantation. Our method uses features generated from Dynamic Contrast-Enhanced Magnetic Resonance Imaging from 4 follow-ups during the first year after the transplant surgery. To deal with missing data, a key mask tensor exploiting the dot product attention mechanism of the transformers is used. Moreover, different contrastive schemes based on cosine similarity distance are proposed to handle the limited amount of available data. Trained on 69 subjects, our best model achieves 96.3% F1 score and 98.9% ROC AUC in the prediction of serum creatinine threshold on a separated test set of 20 subjects. Thus, our experiments highlight the relevance of considering sequential imaging data for this task and therefore in the study of chronic dysfunction mechanisms in renal transplantation, setting the path for future research in this area. Our code is available at https://github.com/leomlck/renal_transplant_imaging.

Keywords: Sequential architectures · missing data · contrastive learning · renal transplant · MRI

1 Introduction

Renal transplantation appears as the most effective solution for end-stage renal disease and highly improves patients' quality of life, mainly by avoiding periodic

dialysis [24]. However, a substantial risk of transplant chronic dysfunction or rejection persists and may lead to graft loss or ultimately the patient death [11]. The genesis of such events takes place in heterogeneous causes, complex phenomena, and results from a gradual decrease in kidney function. In clinical practice, the primary indicator of kidney function is based on blood tests and urine sampling (serum creatinine, creatinine clearance). However, when results are irregular, the gold standard method is needle biopsy, an invasive surgical operation. Thus, the need for a non-invasive alternative that could provide valuable information on transplant function post-transplantation through time is crucial.

Medical imaging plays a significant role in renal transplantation. In [21], diverse imaging modalities have been investigated to assess renal transplant functions in several studies. Moreover, in [17] multiple Magnetic Resonance Imaging (MRI) modalities are used for the unsupervised kidney graft segmentation. Beyond the respective limitations of the several imaging modalities, such as the necessity of radiations or the intrinsic trade-off on resolution, to our knowledge, there are no studies focusing on monitoring the evolution of kidney grafts using imaging data. On the other hand, the recent transformer models [26] offer new directions in processing sequential data. Moreover, recent advances in self-supervised learning [25] enable the training of powerful deep learning representations with a limited amount of data. Renal transplantation datasets usually belong to this case, making the use of such methods the way to move forward. Our study is among the first that explore such methods for renal transplantation, solving challenging clinical questions.

In this work, we propose a method to forecast the renal transplant function through the serum creatinine prediction from follow-up exams of Dynamic Contrast-Enhanced (DCE) MRI data post-transplantation. The main contributions of this work are twofold. First, we propose the use of contrastive schemes, generating informative manifolds of DCE MRI exams of patients undergoing renal transplantation. Different self-supervised and weakly-supervised clinical pertinent tasks are explored to generate relevant features using the cosine similarity. Secondly, we introduce a transformer-based architecture for forecasting serum creatinine score, while proposing a tailored method to deal with missing data. In particular, our method is using a key mask tensor that highlights the missing data and does not take them into account for the training of the sequential architecture. Such a design is very robust with respect to the position and number of missing data, while it provides better performance than other popular data imputation strategies. To the best of our knowledge, our study is among the first that propose a novel, robust, and clinically relevant framework for forecasting serum creatinine directly from imaging data.

2 Related Work

Several medical imaging approaches investigated the diagnosis of renal transplant dysfunction. Recent studies focused on detecting specific events such as renal fibrosis [18] or acute rejection [14]. In [22], multi-modal MRI and clinical data are

explored to assess renal allograft status at the time of the different exams. Most of those approaches seek to, indirectly through related events or directly through complex automated systems, non-invasively retrieve structural, functional, and molecular information to diagnose chronic kidney disease [1].

When it comes to real clinical settings, missing data is one of the most important issues during data curation. Handling of missing data has been thoroughly studied by data imputation methods, which mainly propose approaches to fill the missing data as a pre-processing step to some downstream task [16]. Beyond simple statistical approaches such as sampling the mean or median of available data, methods can be categorized into two groups: discriminative and generative approaches. The former is mainly developed for structural data (discrete or continuous) with methods such as structured prediction [13]. On the other hand, generative approaches include expectation-maximization algorithms [8] or deep learning models such as Generative Adversarial Imputation Nets (GAIN) [29]. Those latest approaches showed very good performance for medical image tasks, as proposed in [5,28]. However, the training of such models usually is subjective to a big amount of data that are not all the time available [12], especially in a clinical setting.

Considering the use of the transformer models, the attention mechanism showed promising results in missing data imputation for structural [27] and trajectory data [2,9]. In particular, the attention mask was used to investigate the robustness of a vanilla encoder-decoder transformer and a Bidirectional Transformer (BERT) model [7] while missing 1 to 6 point's coordinates out of 32 for forecasting the people trajectories. Among all these methods, our method is the first to handle in an efficient and robust way missing data with high dimensionality, tested on sequences with long time dependencies.

3 Method

In this study, we focus on the prediction of serum creatinine from imaging data and in particular DCE MRI, exploring both anatomical and functional information. An overview of our method is presented in Fig. 1.

3.1 Contrastive learning for renal transplant

In this work, we propose two contrastive learning schemes to explore meaningful data representations: (a) a self-supervised scheme, where we learn meaningful features by solving the proxy task of determining if two MRI volumes belong to the same patient, and (b) a weakly-supervised scheme, where we discriminate samples based on the differences in the value of various clinical variables.

Let us denote $(v_1, v_2) \in (\mathbb{R}^{N_x \times N_y \times N_z})^2$ a pair of MRI regions of interest. Each stream $i = 1, 2$ consists of a ResNet model to extract a latent representation from the MRI volumes, which takes v_i as input and outputs features $z_i \in \mathbb{R}^{D_f}$, with $D_f = 512$ for ResNet18. Then, a feature embedding head associates these

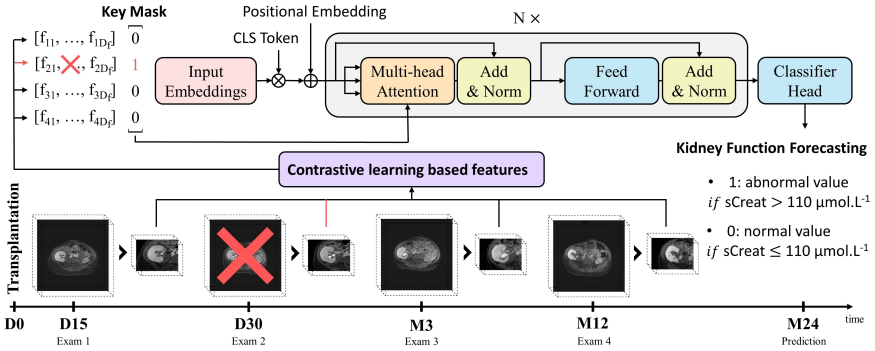

Fig. 1. Overview of the proposed method. Different contrastive schemes are used to represent the different MRIs. These features are used to train a sequential model coupled with a key mask tensor to mark the missing data.

features with the underlying task. This is modeled by a linear layer or a Multi-Layer Perceptron (MLP) mapping the features to $(z'_1, z'_2) \in \mathbb{R}^{D_{fe}}$, with $D_{fe} = 256$.

Self-supervised pre-training. Our first strategy relies on a self-supervised task at the patient level, i.e., we train a model to distinguish if a pair of volumes comes from the same patient or not. $P_j = \{v \in \mathbb{R}^{N_x \times N_y \times N_z} | v \text{ from patient } j\}$ for $j \in [\![1, N_p]\!]$, where N_p denotes the number of patients, the set of available volumes from MRI series for each exam and patient. Then, our proxy task is to discriminate pairs by knowing if they belong or not to the same patient, i.e., $y = 1$ if $\exists j \ (v_1, v_2) \in (P_j)^2$; else $y = 0$.

Weakly-supervised various clinical pre-training. Our second strategy discriminates samples based on the difference of certain clinical variable's value, i.e., $y = 1$ if $\|\text{Var}(v_1) - \text{Var}(v_2)\| < \theta$; else $y = 0$, where $Var(\cdot)$ is a clinicobiological variable and θ a clinically relevant threshold. The clinicobiological variables are suggested by nephrology experts to encode clinical priors and information, as they are significantly linked to graft survival [15]. In this paper, we investigate three variables: (1) the transplant incompatibility, (2) the age of the transplant's donor, and (3) the Glomerular Filtration Rate (GFR) value.

Training Loss. From the embedded features (z'_1, z'_2), the optimization is done by the following cosine embedding loss:

$$\text{CosEmbLoss}(z'_1, z'_2, y) = \begin{cases} 1 - \cos(z'_1, z'_2), & \text{if } y = 1, \\ \max(0, \cos(z'_1, z'_2), & \text{if } y = 0, \end{cases} \tag{1}$$

where *cos* refers to the cosine similarity. This loss enforces the model to build relevant features that express adequately the kidney transplant imaging and define the way to create strategies to label y each pair.

Training Scheme and Curriculum Learning. Since the dimensionality of our data is very high and the tasks we investigate are very challenging, we apply curriculum learning to facilitate the training process. In particular, for the self-supervised task at the patient level, pairs from the same exam of each patient are enabled in the beginning until half of the training, while they are discarded in the second half.

For the weakly-supervised task based on a clinicobiological variables, the perplexity of the task is determined by the thresholds θ. More specifically, the training labels are adjusted every e_i epochs following the rule: $y = 1$ if $|\text{Var}(v1) - \text{Var}(v2)| < \theta_{i,1}$; $y = 0$ if $|\text{Var}(v_1) - \text{Var}(v_2)| > \theta_{i,2}$; else discard the pair (v_1, v_2), where $\theta_{i,1}$, $\theta_{i,2}$ are set in the image of $Var(\cdot)$ satisfying $\forall i$ (1) $\theta_{i,1} \leq \theta_{i,2}$; (2) $\theta_{i+1,1} \leq \theta_{i,1}$; and (3) $\theta_{i,2} \leq \theta_{i+1,2}$. Our loss enforces the feature pairs to be near or far in the feature embedding space, depending on the label y. The condition (1) enables to form a grey area between the two cases, while the conditions (2) and (3) strengthen the constraint through epochs on the difference of value $Var(\cdot)$ between the two pairs to be correctly arranged.

3.2 Sequential model architecture

Our forecasting model takes as input $T = 4$ features $z \in \mathbb{R}^{D_f}$ corresponding to the different follow-ups and relies on a transformer encoder architecture [26]. First, these features are mapped to embeddings of size D_{model} using a linear layer, while a special classification token (CLS) is aggregated in the first position to generate an embedded sequence. Then, the core of the transformer encoder architecture stacks N layers on top of learned positional embeddings added to the embedded sequences. Each layer is first composed of a multi-head self-attention sub-layer, which consists of h heads running in parallel. Each head is based on the scaled dot-product attention. Then, a position-wise fully connected feed-forward sub-layer applies an MLP of hidden dimension D_{model} to each position separately and identically. Finally, to perform the classification task, the CLS token output is fed to a linear layer.

Strategy for missing data. Our proposed strategy to deal with missing data is applied to the scaled dot product operation, core of each multi-head self-attention sub-layer. For simplicity, we consider here a sub-layer with one head, $h = 1$. The operation takes as input the query Q, key K and value V, which are linear projections of the embedded sequences, with d_k, d_k and d_v dimensions, respectively and performs $\text{Attention}(Q, K, V) = \text{softmax}(\frac{(QK^t)}{\sqrt{d_k}})V$. In this work, we build a key mask tensor $m_k \in \mathbb{R}^T$ based on the availability of exams for each patient so that zero attention is given to missing data both during the training and inference times, i.e. $\forall t \in [\![1, T]\!]$ $m_k[t] = -\infty$ if exam t is available else 0. Thus, our mask cancels the attention on missing exams by $\text{Attention}(Q, K, V) = \text{softmax}(\frac{(QK^t)}{\sqrt{d_k}} + M_k)V$ where $M_k = [\![m_k m_k ... m_k]\!] \in \mathbb{R}^{T \times d_k}$. For $h > 1$, keys, values, and queries are linearly projected h times with different, learned linear projections, concatenated, and once again projected after the scaled-dot product.

3.3 Implementation Details

Starting with the contrastive learning, we used data augmentation with horizontal flipping and random affine transformation with a 0.5 probability, as well as random Gaussian blur ($\sigma \in [0, 0.5]$) and random Gaussian noise ($\sigma \in [0, 0.05]$), using TorchIO python library [20]. Having approximately a set of pairs of $\binom{V}{2} = \frac{V(V-1)}{2}$, where V is the number of available volumes, we proposed to fix the training set size to $V_t = 5000$. We decided to fix the number of positive samples, as well as its balance to 25%, and to randomly sample every epoch the remaining from the negative samples.

Concerning the optimization of our models, a 10% dropout has been used for the linear layers of both the contrastive and sequential models. For the contrastive model, the Stochastic Gradient Descent optimizer with a momentum equal to 0.9 was used with a starting learning rate of $1e^{-2}$ following a cosine schedule and preceded by a linear warm-up of 5 epochs. The batch size was set to 20 and the model trained for 60 epochs on 4 NVIDIA Tesla V100 GPU using Pytorch [19]. For the transformer, a binary cross-entropy loss (BCE) was used when binarizing the serum creatinine value using a threshold of $110 \mu mol.L^{-1}$, specified by nephrology experts, as a clinically relevant value to assess normal/abnormal renal transplant function at a specific time point. Adam optimizer was used with a starting learning rate of $1e^{-4}$ following the same learning rate scheduler. The batch size was set to 32 and the model was trained for 30 epochs on 1 NVIDIA Tesla V100 GPU. The architecture's hyperparameters were set by grid search and 10-fold cross-validation, providing $N = 2$, $h = 2$, $D_{model} = 768$.

4 Data

Our study was approved by the Institutional Review Board, which waived the need for patients' consent. The data cohort corresponds to study reference ID-RCB: 2012-A01070-43 and ClinicalTrials.gov identifier: NCT02201537. All the data used in this study were anonymized. Overall, our imaging data are based on DCE MRI series collected from 89 subjects at 4 follow-up exams which took place approximately 15 days (D15), 30 days (D30), 3 months (M3), and 12 months (M12) after the transplant surgery, resulting in respectively 68, 75, 87, and 83 available scans at each follow-up.

The MRI volumes sized $512 \times 512 \times [64 - 88]$ voxels included spacing ranging in $[0.78 - 0.94] \times [0.78 - 0.94] \times [1.9 - 2.5]$ mm. All volumes were cropped around the transplant using an automatic selection of the region of interest in order to reduce dimensionality while no information about the transplant is discarded. Intensity normalization was executed to each volume independently by applying standard normalization, clipping values to $[-5, 5]$ and rescaling linearly to $[0, 1]$.

As a primary indicator of the kidney function assessment, all patients were subject to blood tests regularly a few days before the transplantation to several years after, to measure the serum creatinine level in $\mu mol.L^{-1}$. The serum creatinine target prediction value is calculated as the mean over an interval of two months before and after the prediction date, 2-year post-transplantation (M24).

Table 1. Quantitative evaluation of the proposed method against other methods. sCreat stands for simple statistics from the serum creatinine and Radiomics for predefined radiomics features [10], including shape, intensity, and texture imaging features. We report in format mean(std): Precision (Prec), Recall (Rec), F1 score, and ROC AUC (AUC). **Bold** indicates the top-performing combination.

Method	Features	Validation				Test			
		Prec	Rec	F1	AUC	Prec	Rec	F1	AUC
LSTM	**sCreat**	80,5(12,3)	62,9(21,0)	71,1(13,8)	80,4(22,4)	83,3	76,9	80,0	83,5
	Radiomics [10]	86,2(14,9)	73,5(15,5)	77,3(8,2)	80,7(16,0)	90,9	76,9	83,3	84,6
	Imagenet [6]	85,5(15,0)	68,0(17,7)	74,0(12,8)	91,0(10,8)	90,9	76,9	83,3	81,3
	Kinetics [23]	**90,7(9,4)**	74,0(21,5)	78,5(11,3)	**91,4(8,5)**	92,3	92,3	92,3	85,7
	MedicalNet [3]	86,5(13,9)	78,5(21,2)	79,8(13,2)	82,7(18,8)	57,1	61,5	59,3	41,8
	SimCLR [4]	79,8(15,9)	86,5(24,1)	80,9(17,2)	91,8(13,7)	72,2	**100,0**	83,9	64,8
	Proposed GFR	82,8(9,6)	**95,5(9,1)**	**88,3(7,7)**	88,3(13,1)	86,7	**100,0**	92,9	**98,9**
Transformer	**sCreat**	79,0(28,7)	60,2(31,1)	65,4(29,3)	71,6(24,2)	81,3	**100,0**	89,7	86,8
	Radiomics [10]	81,3(15,7)	66,0(28,6)	69,1(20,2)	65,3(30,5)	90,9	76,9	83,3	91,2
	Imagenet [6]	58,4(22,4)	76,5(34,8)	65,8(27,5)	45,5(21,6)	65,0	**100,0**	78,8	58,2
	Kinetics [23]	53,2(35,8)	66,0(44,8)	58,3(38,9)	64,0(19,7)	65,0	**100,0**	78,8	83,5
	MedicalNet [3]	65,5(27,9)	58,0(33,2)	58,3(28,3)	64,8(19,6)	75,0	46,2	57,1	50,6
	SimCLR [4]	58,9(30,9)	75,5(38,7)	65,6(33,2)	64,8(23,5)	68,4	**100,0**	81,3	72,5
	Proposed GFR	86,3(20,9)	71,5(22,7)	77,4(20,6)	79,7(20,7)	**92,9**	**100,0**	**96,3**	**98,9**

5 Experiments & Analysis

To evaluate the performance of our proposed method and compare it with other strategies for the forecasting of serum creatinine, four evaluation metrics are used: recall, precision, F1 score, and the area under the receiver operating characteristic curve (ROC AUC). A testing set of 20 patients is separated from the train set and used to validate the performance of our models. We perform a 10-fold cross-validation (CV) on the train set (69 patients) and report the mean (standard deviation) scores in % for each fold. During CV, the model reaching the minimum loss is saved, and an ensemble approach is used to make the final prediction on the test set from models, which reach more than 50% ROC AUC out of the 10 folds.

We compare our sequential model to an LSTM model, which is a commonly used architecture for sequential data, and which architecture was set using the same approach as our main model, resulting in 2 LSTM cells and a hidden size of 768. Additional sets of feature representations were used to compare the significance of our approach. First simple statistics from the serum creatinine captured from the available blood test results between each follow-up (number of points, mean, median, standard deviation, minimum, maximum) are calculated and used as input to the models. Second, a set of predefined radiomics features [10] are obtained from the segmentation of the kidney transplant following the unsupervised method presented in [17]. Finally, we investigate generating MRI features from SimCLR [4] contrastive scheme, while we report the performance of different transfer-learning approaches, pre-trained on ImageNet [6] by duplicating the weights to 3D, Kinetics [23], and medical image datasets MedicalNet [3].

Quantitative results for all the methods are reported in Table 1. Our proposed approach outperforms the rest of the methods for the test. Both LSTMs and transformers architectures report good performances, with only a few models reporting performance lower than 60% on every metric. Interestingly, our method outperforms the sCreat model which models directly the serum creatinine level. Moreover, our GFR contrastive-based features report the best performance among all the other features for both LSTMs and transformer formulations. The rest of the pre-training performances are summarised in the supplementary materials. Limitations appear as our model seems to misclassify cases where the patient's serum creatinine is stable and close to the used threshold, during the first two years post-transplantation.

5.1 Ablation study for missing data strategies

The proposed key mask padding approach for handling missing data is specific to the attention mechanism, hence the transformer model. Thus, we investigate 3 other missing data strategies applicable to both the transformer and LSTM model: (1) filling with zeros strategy (None), (2) filling with the nearest available exam (N.A.), and (3) taking the mean for intermediate exams and fill for first and last (M.+N.A.). Results presented in Table 2 are obtained with the best performing imaging features (proposed using the GFR value).

Our proposed approach to handling missing data reports the best precision, recall, and F1 score and the second-best ROC AUC on the test set. Overall, the different strategies report better performance on transformer based architectures than the LSTMs ones indicating the interest in using such models for this task. Moreover, the M.+N.A. strategy reports a lower precision rate for both LSTM and our sequential model, affirming the difficulty to interpolate imaging features. Both None and N.A. strategies appear to report competitive results, lower however from our proposed.

Table 2. Quantitative evaluation of different strategies for missing data. With none we denote the filling with zero strategy, N.A. the filling with the nearest neighbor exam, and with M.+N.A. the filling with the mean and nearest neighbor exam. **Bold** indicates the top performing combination.

Method	Strategy	Validation				Test			
		Prec	Rec	F1	AUC	Prec.	Rec	F1	AUC
LSTM	None	$80,5(11,5)$	$81,0(14,3)$	$80,0(9,6)$	$73,6(16,9)$	$86,7$	**100,0**	$92,9$	$98,9$
	N.A.	$82,8(9,6)$	**95,5(9,1)**	**88,3(7,7)**	**88,3(13,1)**	$86,7$	**100,0**	$92,9$	$98,9$
	M.+N.A.	$81,1(10,8)$	$93,5(10,0)$	$86,1(6,8)$	$84,2(11,0)$	$81,3$	**100,0**	$89,6$	$96,7$
Transformer	None	$86,2(12,9)$	$78,5(23,2)$	$79,7(18,2)$	$71,5(25,3)$	$92,3$	$92,3$	$92,3$	$98,9$
	N.A.	$88,8(20,6)$	$75,5(21,5)$	$81,3(20,8)$	$80,5(22,2)$	$92,3$	$92,3$	$92,3$	$96,7$
	M.+N.A.	**90,5(12,3)**	$73,5(17,3)$	$80,0(12,7)$	$80,0(18,3)$	$76,5$	**100,0**	$86,7$	**100,0**
	Proposed	$86,3(20,9)$	$71,5(22,7)$	$77,4(20,6)$	$79,7(20,7)$	**92,9**	**100,0**	**96,3**	$98,9$

6 Conclusion

This study proposes a novel transformer based architecture tailored to deal with missing data for the challenging task of serum creatinine prediction 2 years post-transplantation using imaging modalities. First, we show the significant use of contrastive learning schemes for this task. Our trained representations outperform common transfer learning and contrastive approaches. Then, a transformer encoder architecture enables to input the sequential features data per follow-up in order to forecast the renal transplant function, including a custom method to handle missing data. Our strategy performs better than other commonly used data imputation techniques. Those promising results encourage the use of medical imaging over time to assist clinical practice for fast and robust monitoring of kidney transplants.

Acknowledgements This work was performed HPC resources from the "Mésocentre" computing center of CentraleSupélec and École Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France.

References

1. Alnazer, I., Bourdon, P., Urruty, T., Falou, O., Khalil, M., Shahin, A., Fernandez-Maloigne, C.: Recent advances in medical image processing for the evaluation of chronic kidney disease. Med. Image Anal. **69**, 101960 (2021)
2. Becker, S., Hug, R., Huebner, W., Arens, M., Morris, B.T.: Missformer: (in-)attention-based handling of missing observations for trajectory filtering and prediction. In: Advances in Visual Computing: 16th International Symposium (ISVC). p. 521–533. Springer International Publishing (2021)
3. Chen, S., Ma, K., Zheng, Y.: Med3D: Transfer learning for 3D medical image analysis. arXiv preprint, arXiv:1904.00625 (2019)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: 37th International Conference on Machine Learning (ICML). vol. 119, pp. 1597–1607. PMLR (2020)
5. Dalca, A.V., Bouman, K.L., Freeman, W.T., Rost, N.S., Sabuncu, M.R., Golland, P.: Medical image imputation from image collections. IEEE Trans. Med. Imaging **38**, 504–514 (2019)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 248–255. IEEE (2009)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL). vol. 1, pp. 4171–4186. Association for Computational Linguistics (2019)
8. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Comput. Appl. **19**(2), 263–282 (2010)
9. Giuliari, F., Hasan, I., Cristani, M., Galasso, F.: Transformer networks for trajectory forecasting. In: 25th International Conference on Pattern Recognition (ICPR). pp. 10335–10342. IEEE (2021)

10. van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.W.L.: Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res. **77**(21), e104–e107 (2017)
11. Hariharan, S., Israni, A.K., Danovitch, G.: Long-term survival after kidney transplantation. N. Engl. J. Med. **385**(8), 729–743 (2021)
12. Kazeminia, S., Baur, C., Kuijper, A., van Ginneken, B., Navab, N., Albarqouni, S., Mukhopadhyay, A.: Gans for medical image analysis. Artif. Intell. Med. **109**, 101938 (2020)
13. Keshavan, R.H., Montanari, A., Oh, S.: Matrix completion from a few entries. IEEE Trans. Inf. Theory **56**(6), 2980–2998 (2010)
14. Khalifa, F., Beache, G.M., El-Ghar, M.A., El-Diasty, T., Gimel'farb, G., Kong, M., El-Baz, A.: Dynamic contrast-enhanced mri-based early detection of acute renal transplant rejection. IEEE Trans. Med. Imaging **32**(10), 1910–1927 (2013)
15. Loupy, A., Aubert, O., Orandi, B.J., Naesens, M., Bouatou, Y., Raynaud, M., Divard, G., Jackson, A.M., Viglietti, D., Giral, M., Kamar, N., Thaunat, O., Morelon, E., Delahousse, M., Kuypers, D., Hertig, A., Rondeau, E., Bailly, E., Eskandary, F., Böhmig, G., Gupta, G., Glotz, D., Legendre, C., Montgomery, R.A., Stegall, M.D., Empana, J.P., Jouven, X., Segev, D.L., Lefaucheur, C.: Prediction system for risk of allograft loss in patients receiving kidney transplants: international derivation and validation study. BMJ **366**, l4923 (2019)
16. Mackinnon, A.: The use and reporting of multiple imputation in medical research – a review. J. Intern. Med. **268**(6), 586–593 (2010)
17. Milecki, L., Bodard, S., Correas, J.M., Timsit, M.O., Vakalopoulou, M.: 3D unsupervised kidney graft segmentation based on deep learning and multi-sequence MRI. In: 18th International Symposium on Biomedical Imaging (ISBI). pp. 1781–1785. IEEE (2021)
18. Orlacchio, A., Chegai, F., Del Giudice, C., Anselmo, A., Iaria, G., Palmieri, G., Di Caprera, E., Tosti, D., Costanzo, E., Tisone, G., Simonetti, G.: Kidney transplant: Usefulness of real-time elastography (rte) in the diagnosis of graft interstitial fibrosis. Ultrasound Med. Biol. **40**(11), 2564–2572 (2014)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
20. Pérez-García, F., Sparks, R., Ourselin, S.: Torchio: A python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. **208**, 106236 (2021)
21. Sharfuddin, A.: Renal relevant radiology: imaging in kidney transplantation. Clin. J. Am. Soc. Nephrol. **9**(2), 416–429 (2014)
22. Shehata, M., Ghazal, M., Khalifeh, H.A., Khalil, A., Shalaby, A., Dwyer, A.C., Bakr, A.M., Keynton, R., El-Baz, A.: A deep learning-based cad system for renal allograft assessment: Diffusion, bold, and clinical biomarkers. In: International Conference on Image Processing (ICIP). pp. 355–359. IEEE (2020)
23. Smaira, L., Carreira, J., Noland, E., Clancy, E., Wu, A., Zisserman, A.: A short note on the kinetics-700-2020 human action dataset. arXiv preprint, arXiv:2010.10864 (2020)
24. Suthanthiran, M., Strom, T.B.: Renal transplantation. N. Engl. J. Med. **331**(6), 365–376 (1994)

25. Taleb, A., Loetzsch, W., Danz, N., Severin, J., Gaertner, T., Bergner, B., Lippert, C.: 3D self-supervised methods for medical imaging. In: Advances in Neural Information Processing Systems. vol. 33, pp. 18158–18172. Curran Associates, Inc. (2020)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
27. Wu, R., Zhang, A., Ilyas, I., Rekatsinas, T.: Attention-based learning for missing data imputation in holoclean. In: Conference on Machine Learning and Systems (MLsys). vol. 2, pp. 307–325 (2020)
28. Xia, Y., Zhang, L., Ravikumar, N., Attar, R., Piechnik, S.K., Neubauer, S., Petersen, S.E., Frangi, A.F.: Recovering from missing data in population imaging – cardiac mr image imputation via conditional generative adversarial nets. Med. Image Anal. **67**, 101812 (2021)
29. Yoon, J., Jordon, J., van der Schaar, M.: GAIN: Missing data imputation using generative adversarial nets. In: 35th International Conference on Machine Learning (ICML). vol. 80, pp. 5689–5698. PMLR (2018)