Automated prognosis of renal function decline in ADPKD patients using deep learning

Anish Raj [a,b,*], **Fabian Tollens** [c], **Anna Caroli** [d], **Dominik Nörenberg** [c], **Frank G. Zöllner** [a,b]

[a] Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim 68167, Baden Württemberg, Germany
[b] Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim 68167, Baden Württemberg, Germany
[c] Department of Radiology and Nuclear Medicine, University Medical Centre Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim 68167, Baden Württemberg, Germany
[d] Bioengineering Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Ranica, BG 24020, Italy

Received 11 April 2023; accepted 1 August 2023

Abstract

An accurate prognosis of renal function decline in Autosomal Dominant Polycystic Kidney Disease (ADPKD) is crucial for early intervention. Current biomarkers used are height-adjusted total kidney volume (HtTKV), estimated glomerular filtration rate (eGFR), and patient age. However, manually measuring kidney volume is time-consuming and subject to observer variability. Additionally, incorporating automatically generated features from kidney MRI images, along with conventional biomarkers, can enhance prognostic improvement. To address these issues, we developed two deep-learning algorithms. Firstly, an automated kidney volume segmentation model accurately calculates HtTKV. Secondly, we utilize segmented kidney volumes, predicted HtTKV, age, and baseline eGFR to predict chronic kidney disease (CKD) stages >=3A, >=3B, and a 30% decline in eGFR after 8 years from the baseline visit. Our approach combines a convolutional neural network (CNN) and a multi-layer perceptron (MLP). Our study included 135 subjects and the AUC scores obtained were 0.96, 0.96, and 0.95 for CKD stages >=3A, >=3B, and a 30% decline in eGFR, respectively. Furthermore, our algorithm achieved a Pearson correlation coefficient of 0.81 between predicted and measured eGFR decline. We extended our approach to predict distinct CKD stages after eight years with an AUC of 0.97. The proposed approach has the potential to enhance monitoring and facilitate prognosis in ADPKD patients, even in the early disease stages.

Keywords: Autosomal dominant polycystic kidney disease; Deep learning; Total kidney volume; Chronic kidney disease; Image classification; Regression

1 Introduction

Autosomal dominant polycystic kidney disease (ADPKD), due to the growth of cysts and therefore, degeneration of renal parenchyma leads to end-stage renal disease (ESRD) and renal failure [1]. It affects up to 12 million people worldwide. ADPKD accounts for up to 10% of patients with ESRD [1]. Given these numbers, early treatment of the

* Corresponding author: Anish Raj, M.Sc., Computer Assisted Clinical Medicine, Medical Faculty Mannheim, Heidelberg University, Mannheim 68167, Baden Württemberg, Germany.
E-mail: Anish.Raj@medma.uni-heidelberg.de (A. Raj).

https://doi.org/10.1016/j.zemedi.2023.08.001
www.elsevier.com/locate/zemedi

disease when the renal parenchyma is still preserved is warranted. However, disease progression monitoring is difficult since kidney function may remain normal for several decades and is therefore not informative in the earliest stages of the disease. Clinical, genetic, environmental, epigenetic, and radiologic factors have been studied as predictors of progression to kidney failure in ADPKD [2,3]. On the other hand, kidney volume has been shown to be a promising marker for disease progression [4,5]. It is also recognized by the Food and Drug Administration (FDA) as a candidate for a prognostic biomarker for ADPKD progression [6] and used within recent clinical phase 3 studies (TEMPO 3/4 trial, primary outcome measure; the annual rate of change in TKV over time) [7,8].

Magnetic resonance imaging (MRI) is recognized as an important tool to monitor the progression of ADPKD, and the total kidney volume (TKV) or the height-adjusted total kidney volume (HtTKV) has been shown to predict renal function decline in ADPKD patients [9,2].

Thereby, complex interactions of different prognostic factors based on clinical, genetic, environmental, and radiological information determine the number of kidney cysts and their growth rates, which affect the total kidney volume (TKV)[2] and ultimately, renal function decline [10]. Based on this, prognostic models like the Mayo imaging classification tool [9] have been developed to stratify ADPKD patients into classes and predict disease progression. A multiple linear regression model is generally employed for this task. However, Kline et al. [11] have shown that using the conventional biomarkers (HtTKV, age, and eGFR) may not be sufficient for accurate predictions. They reported that for predicting eGFR decline after 8 years, a Pearson correlation coefficient 'r' of only 0.51 could be achieved in classifying whether a patient will reach CKD stage 3A or not (eGFR $<$ 60 ml/min/1.73 m^2), reach 3B or not (eGFR $<$ 45 ml/min/1.73 m^2), and reach a 30% decline in eGFR or not. They improved this prediction model by further incorporating texture features (repeating patterns of local image intensity variations that provide information regarding the spatial arrangement of image intensities) from T2-weighted MRI images. The texture features are extracted from manually segmented kidneys and consist of energy, entropy, and correlation values. By adding texture features to conventional biomarkers, the resulting Pearson's r reaches -0.70. However, this approach, requiring manual segmentation of the kidneys, is time-consuming and observer-dependent. Moreover, automated kidney segmentation approaches are steadily proposed and are emerging into different applications in renal MRI [12,13]. Considering this, we created a fully automated system that can segment the kidneys automatically and can make an accurate prognosis. More in detail, we first develop a deep learning model that segments kidneys from T2-weighted MRI, enabling calculation of HtTKV. We then extract image features from the segmented kidneys and combine them with the conventional biomarkers (HTKV, age, eGFR) to predict renal function decline after 8 years, using data from 135 ADPKD patients with normal kidney function at baseline. Specifically, we classify each patient into a distinct CKD stage (CKD stages 1, 2, 3A, 3B, and 4) and predict percent change in the eGFR.

2 Methods

2.1 Patient data

The patient data were acquired from the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK), National Institutes of Health, USA and were recorded in the Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) study [4]. The dataset consists of 241 patients in total. However, for our work, we selected patients who underwent T2-weighted MR imaging and had

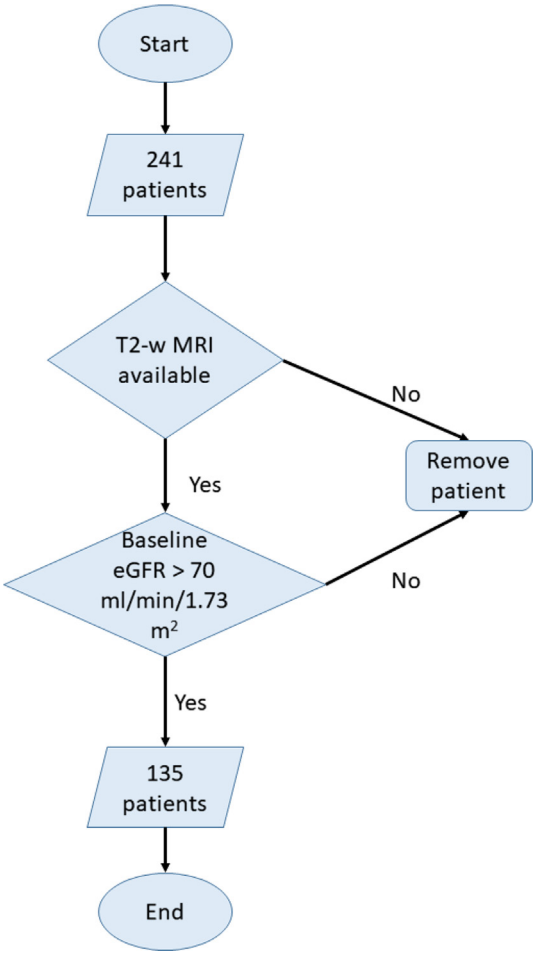

Figure 1. Flow chart for patient selection based on criteria from [9,11].

eGFR values $>$ 70 ml/min/1.73 m^2 at the baseline visit as depicted in Fig. 1. These criteria were selected in accordance with [11,9]. The resulting dataset contains information from 135 patients. Table 1 list the number of patients' CKD stages at baseline and after 8 years alongside demographic data, eGFR, and kidney volumes.

To allow for a comparison to the SOTA method [11], the patients were regrouped into two groups based on the presence or absence of each of the following conditions at 8 years after baseline: reaching CKD stage 3A (eGFR $<$ 60 ml/min/1.73 m^2), reaching CKD stage 3B (eGFR $<$ 45 ml/min/1.73 m^2), and 30% decline in eGFR. As depicted in Table 2, our data resembles a similar class distribution as in the work of Kline et al. [11].

2.2 Data pre-processing

We first normalized the renal MRIs, predicted HtTKV (ml/m) (obtained as per Section 2.4.1), age (years), and eGFR (ml/min/1.73 m^2) at the baseline visit using the Z-score normalization Eq. (1),

$$\hat{I} = \frac{I - \mu(I)}{\sigma(I)}, \tag{1}$$

where I and $\hat{I}$ are the original and normalized image/feature, respectively. The T2-weighted volumes were recorded with a size of 256 $\times$ 256 pixels and 12–30 slices [4]. However, the network for prognosis requires uniformly sized image volumes as inputs. Hence, we center-cropped and/or padded (as required) the image volumes to the size of 224 $\times$ 224 $\times$ 16 voxel. We ensure that complete kidney volumes were present within this reshaped volume.

2.3 Image annotation

The annotations are used from a previous study on kidney segmentation by Raj et al. [15]. In that study, two physicians independently performed annotation on coronal MR images. The physicians had an experience of one and three years, respectively. Furthermore, they were overseen by an expert abdominal radiologist with nine years of experience. The annotations have the mean inter-user agreement of 0.91 $\pm$ 0.06 (Dice) and a coefficient of variation of 0.07.

Table 1
Demographic, clinical and radiological data of the 135 ADPKD patients included in the study. Data are shown as mean $\pm$ SD or Number (%).

	Baseline	8-year Follow-up
TKV (ml)	974±585	1531±1135
HtTKV (ml/m)	559±322	880±615
eGFR (ml/min/1.73 m^2)	96±25	76±28
CKD stage 1	66	46
CKD stage 2	69	44
CKD stage 3A	-	24
CKD stage 3B	-	14
CKD stage 4	-	7
$>$ 30% eGFR decline	-	48
Gender (M/F)	58/77	-
Age (years)	32±9	-

2.4 Deep learning models

We implement two deep learning models for a fully automatic prognosis. We first used a segmentation network called attention U-Net (Fig. 2) [16] to extract kidneys from MRI volumes. Thereafter, we use the segmented kidney volumes as input to our prognosis network (Fig. 3)) to make the final prognosis. We implement this two-step process so that we can first derive HtTKV automatically and then use the output from the first network and feed it to the second network for the classification/regression task.

2.4.1 Segmentation network

The first network is based on previous work utilizing the attention U-Net for renal segmentation in ADPKD and TKV estimation from T1-weighted MRI [15]. Briefly, this network is composed of a U-shaped encoder-decoder architecture [14] and modified with attention gates [16], which help focus the network on relevant image regions. This is achieved by using higher-level features as a guide to suppress noise and irrelevant features in the lower-level features. The network architecture is illustrated in Fig. 2.

The network was re-trained on 100 T2-weighted volumes of our dataset for which manually segmented kidneys were available. Kidney annotations were done similarly as described in [15]. To obtain kidney segmentations for the remaining patients without manual annotation, the trained network was applied. All automatically segmented kidneys from the network are then used to calculate the HtTKV. The HtTKV can be calculated by multiplying the number of segmented kidney voxels by the voxel volume (in mm^3) and then dividing it by 1000 and the height of the patient (in m) to get the volume in ml/m.

Table 2
Ratio of the positive number of samples to the total number of samples in our dataset as compared to the one in state-of-the-art (SOTA) approach from Kline et al. [11]. In all three criteria, we have more imbalance as compared to the work from Kline et al. [11].

Criteria	Current study	SOTA method
Reached CKD stage 3A	0.333	0.360
Reached CKD stage 3B	0.155	0.180
30% eGFR decline	0.355	0.385

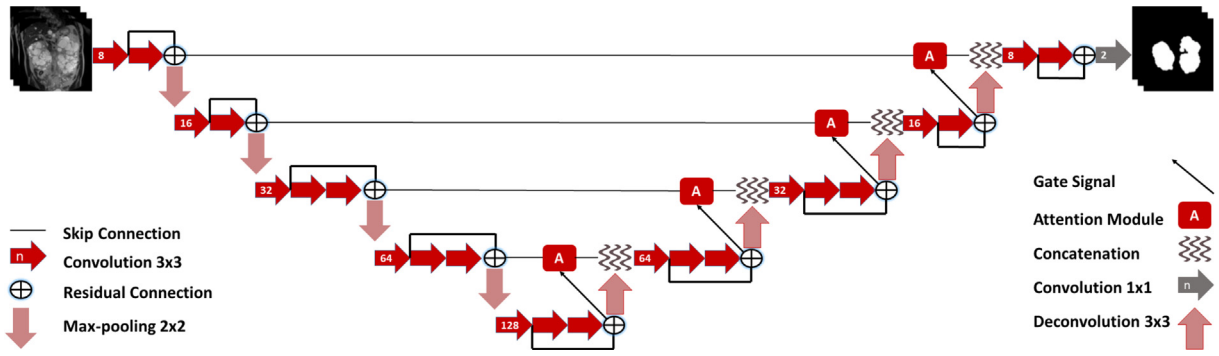

Figure 2. Segmentation network: 2D Attention U-Net with attention modules. The network is U-Net based [14] with attention gates that help the network focus on relevant image regions, e.g., kidneys. The network is used to segment kidneys from patient MRI volumes (baseline visit). While training, the input to the network is a patch size of 128 × 128. During inference, each slice of size 256 × 256 is used as input. The number of filters used in each block is written under the corresponding convolution block. The segmented kidneys are then used to calculate the HtTKV of the patient. Adapted from [15,16].

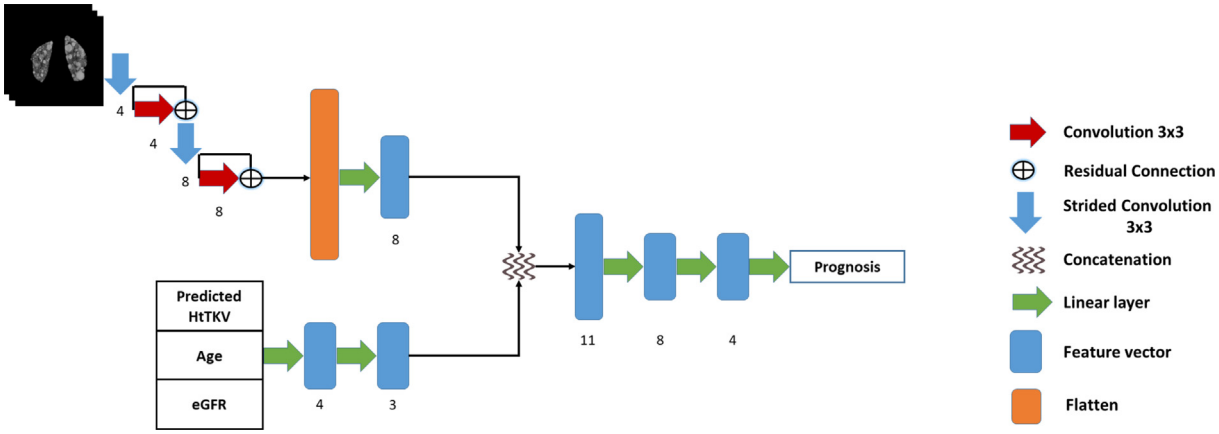

Figure 3. Our proposed prognosis network architecture for classification/regression prognosis. The first input is patient MRI volume (baseline visit) of size 224 × 224 × 16 that contains only segmented kidneys obtained from the segmentation network. The convolution layers are used to extract features from the volumes. The number of filters in each feature vector is written under the corresponding vector. The second input consists of a vector of predicted HtTKV, age, and eGFR at the baseline visit. Here, an MLP extracts features. Then, the features from both inputs are concatenated and used to make the final prognosis. Every convolution, strided convolution, and linear layer is followed by instance normalization and PReLU activation [17]. The strided convolutions have a stride of 4. There is no final activation for the regression task, but for classification, we use sigmoid.

2.4.2 Prognosis network

Our proposed neural network architecture for predicting the final prognosis comprises two components: a Convolutional Neural Network (CNN) and a Multi-Layer Perceptron (MLP). The CNN is employed to extract comprehensive image features from kidney MRI volumes. This CNN component consists of convolutional and strided convolutional layers. The strided convolutions have a stride of 4, resulting in downsampling the spatial dimensions of the image/features by a factor of 4 during each operation. The decision to utilize a shallow CNN is motivated by our limited dataset. Deeper networks typically require larger datasets to achieve robust performance. By employing a shallow network with a stride of 4, we can aggressively downsample the spatial information from the kidneys and extract features without resorting to a deeper network.

The MLP component comprises linear layers that generate one-dimensional feature vectors. We opt for linear layers as they are well-suited for processing and extracting information from structured or tabular data, such as biomarker values. The CNN component takes image volumes as input, while the MLP component accepts a vector composed of

predicted HtTKV, age, and eGFR from the baseline visit. The features extracted from the CNN and MLP are concatenated, and a final MLP block is employed to make a prognosis. The concatenation of features from the CNN and MLP enables the fusion of information derived from kidney MRI volumes and conventional biomarkers.

To enhance the network's performance, we incorporate instance normalization layers [18] and PReLU activation functions [17] after each convolutional, strided convolutional, and linear layer (excluding the last linear layer). We opt for instance normalization instead of batch normalization due to the utilization of a small batch size resulting from our limited dataset. Batch normalization can introduce noise when applied to small batch sizes, thus, to mitigate this issue, we implement instance normalization in our network.

Depending upon the task type, the last layer is followed by the softmax activation in the case of the classification task or no activation in the case of the regression task. The complete network architecture is depicted in Fig. 3.

The prognosis after 8 years is divided into two tasks:

1. Multi-class classification: here, we classify each patient under a distinct CKD stage category after 8 years (CKD stages 1, 2, 3A, 3B, and 4).
2. Regression: percent change in the eGFR.

As illustrated in Fig. 3, we feed the segmented kidneys to a small convolutional neural network. Simultaneously, we use the predicted HtTKV, age, and eGFR at the baseline visit as input to a shallow multi-layer perceptron. Finally, we combine the features from kidney volumes and the MLP inputs and run them through a final MLP to make our prognosis.

For a proof-of-principle and to compare our approach to the SOTA method, we also considered the following task:

- Binary classifications: whether a patient will reach CKD stage 3A or not (eGFR $<$ 60 ml/min/1.73 m^2), reach 3B or not (eGFR $<$ 45 ml/min/1.73 m^2), and reach a 30% decline in eGFR or not.

Therefore, the last layer of the network is followed by a sigmoid function for each binary classification task.

2.5 Network implementations & training

2.5.1 Segmentation network

The segmentation network for kidney segmentation is trained with a batch size of 16 and a patch size of 128. We used the adam optimizer [20] with a learning rate of 0.001. The loss function is cosine loss [15]. As the activation function, we employ exponential linear units (elus) [21] with batch-normalization, dropout (probability = 0.01) and L2 normalization (10^{-7}). Furthermore, we perform a 5-fold cross-validation [19] (as depicted in Fig. 4) and and train for a minimum of 20 epochs. The train:validation:test split was 70:10:20 subjects in each fold without shuffling. We employ early stopping to avoid overfitting and select the network with the highest dice score (validation data) for testing.

2.5.2 Prognosis network

The prognosis network was trained for 30 epochs with 5-fold stratified cross-validation. The train:validation:test split was 95:13:27 subjects in each fold without shuffling. The batch size and learning rate were 8 and 0.001, respectively. For classification tasks, we used weighted cross-entropy (CE) loss function and area under (AUC) the receiver operating characteristic curve (ROC) to select the best network. However, for the task of classifying distinct CKD stages, we use the f1-score to select the best network. We further use a weighted random sampling strategy to deal with the class imbalance in the classification of distinct CKD stages. Here, classes with less number of samples are weighted higher and accordingly sampled more often with replacement during training. Meanwhile, for regression, we implement mean squared error (MSE) loss and use the mean absolute error (MAE) score to choose the best network[22].

2.6 Evaluation

2.6.1 Segmentation evaluation

We evaluate the performance of the segmentation network using the dice (DSC) score and mean surface symmet-

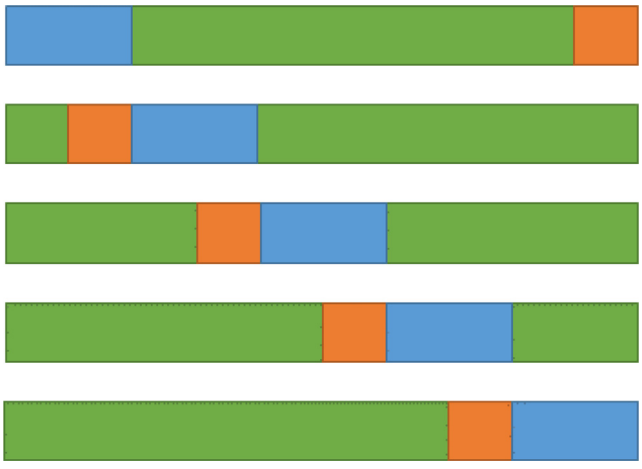

Figure 4. 5-fold cross-validation scheme for training networks [19] The networks are trained 5 times with different training, validation and test sets in each fold. Here, the training, validation and test sets are represented by green, orange and blue bars, respectively. Cross-validation is useful when there is limited dataset, and it helps evaluate the algorithms over a complete dataset.

ric distance (MSSD). First, we used the DSC similarity coefficient to assess the overlap between the ground truth Y and the segmentation $\widehat{Y}$,

$$DSC\left(\widehat{Y}, Y\right) = \frac{2\,|\widehat{Y} \cap Y|}{|\widehat{Y}| + |Y|} \tag{2}$$

Then, we implemented the MSSD (in mm) that is more perceptive to alignment and shape:

$$MSSD\left(\widehat{Y}, Y\right) = \frac{\sum_{\hat{y}\in\widehat{Y}} \min d(\hat{y}, Y) + \sum_{y\in Y} \min d\left(y, \widehat{Y}\right)}{|\widehat{Y}| + |Y|} \tag{3}$$

2.6.2 Prognosis evaluation

To evaluate and compare the results, we used the weighted f1-score (since we have a high class imbalance for CKD stage 4) and the AUC of the ROC for classification outputs. The weighted f1-score is defined as,

$$Weighted\, f1 = \sum_{i=1}^{n=5} Support_i \times f1_i, \tag{4}$$

Where Support_i is the support proportion of each class i, given by,

$$Support_i = \frac{\text{Number of instances in class } i}{\text{Total number of instances}}, \tag{5}$$

and precision, recall, and f1-score are given by

$$Precision = \frac{TP}{TP + FP}, \tag{6}$$

$$Recall = \frac{TP}{TP + FN}, \tag{7}$$

$$f1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{8}$$

The TP, FP, and FN are true positives, false positives, and false negatives, respectively. In addition, we used Pearson's correlation coefficient and Bland–Altman plots for evaluating regression results. Furthermore, for comparison, we employed the Mayo imaging classification tool [9,23] to predict eGFR values at 8 years after baseline, using the baseline HtTKV.

Differences between the Mayo classification tool and our prediction network for eGFR was assessed by a two-sided t-test. Thereby, we compared the error in eGFR between the predicted values and the ground truth at 8 years. The null hypothesis that both are equal is rejected at $p < 0.05$.

3 Results

3.1 Kidney segmentation

The segmentation network is able to segment kidney volumes at high accuracy. After post-processing with the largest connected components, the DSC and MSSD attain the values of 0.909± 0.069 and 0.721 ± 1.484 mm, respectively (see Table 3). Fig. 7 shows three examples of kidney segmentation from the proposed segmentation network.

We then applied the trained segmentation network to the remaining 35 datasets for which no ground truth annotation were available. Visual inspection of the obtained segmentations also showed good accuracy (cf. an example case in Fig. 7, last row).

Inspecting the predicted HtTKV versus ground truth HtTKV a Pearson's correlation coefficient of 0.98 could be obtained (see Fig. 5). In Fig. 5, the regression line has a coefficient of regression r^2 value of 0.96. Furthermore, the mean percent difference between the predicted and the ground truth HtTKV is 13.47 ± 13.70 %. Also, the Bland–Altman plot (see Fig. 6) supports this finding depicting a bias of 66 ml/m.

3.2 Prognosis network

Our proposed approach was tested twofold. First, as a proof-of-concept, we predicted whether a patient reaches CKD stage 3A, 3B, or a 30% decline in eGFR and compare the results with the SOTA method. Furthermore, we predicted eGFR values after 8 years for our patient data using the Mayo imaging classification tool [9,23]. Second, we predicted distinct classes of CKD after 8 years from baseline.

Table 4 shows the results of our approach compared to the SOTA method.

The AUC for reaching CKD stage 3B is over 0.950 and on par with the corresponding value from the state-of-the-art (SOTA) method (0.960) [11]. However, the AUC of reaching CKD stage 3A is 0.965 and is higher than that of the SOTA method (0.940). Further, the 30% eGFR decline predictions reach an AUC of 0.952 and clearly exceed the results of the SOTA method (0.850).

Moreover, we observe that the precision and recall for each criterion is about 90%, indicating good performance of the classifiers (Table 4).

Fig. 8 depicts the predicted versus ground truth eGFR percent change after 8 years. Here, Pearson's correlation

Table 3
Segmentation results for each of the 5 folds after post-processing using largest connected components.

Fold No.	DSC ↑	MSSD (mm) ↓
1	0.918 ± 0.030	0.367 ± 0.123
2	0.908 ± 0.047	0.845 ± 1.209
3	0.940 ± 0.018	0.332 ± 0.155
4	0.906 ± 0.091	0.688 ± 1.130
5	0.874 ± 0.107	1.410 ± 2.870
Average	0.909 ± 0.069	0.721 ± 1.484

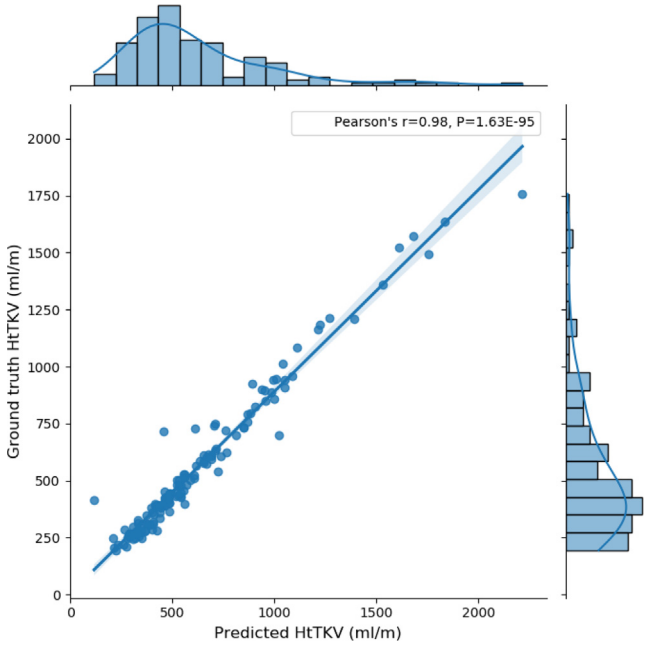

Figure 5. Baseline predicted HtTKV against the ground truth HtTKV. Here, the data represents all 135 patients. The TKV ground truth values for all patients are also provided by the NIDDK CRISP study Chapman et al. [4]. We used these ground truth TKV values to calculate ground truth HtTKV. The predicted HtTKV is obtained from the segmentation network by segmenting kidneys from T2-weighted MRI volumes. Here, we make sure that the predicted data was not utilized during the training of the 5-fold cross-validation model, as shown in Fig. 4. The Pearson correlation coefficient is 0.98.

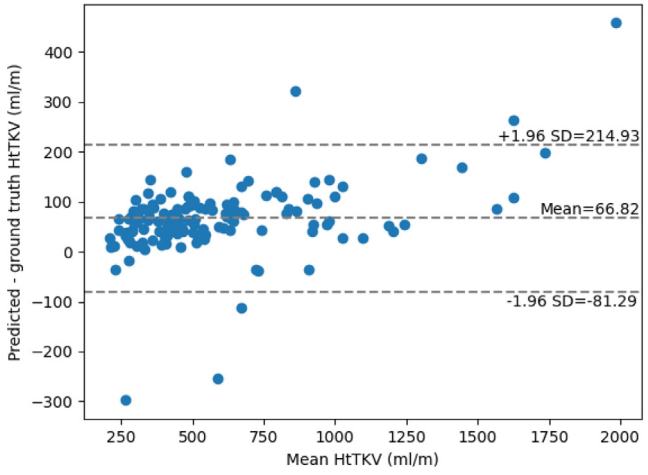

Figure 6. Bland–Altman plot comparing model predicted HtTKV (ml/m) values to the ground truth values.

coefficient of 0.81 is attained (SOTA method = −0.700 [11]). The mean difference between predicted and ground truth eGFR percent change was found to be 1.12 ± 15.58 %. Plotting the respective Bland–Altman plot further supports this observation (see Fig. 9). Most of the data is distributed within the 1.96 standard deviation range with a bias of only 1.12 percent change in eGFR, showing a small overestimation.

Table 4
Classification results of reaching CKD stage 3A, reaching CKD stage 3B and having a 30% eGFR decline after 8 years. Abbreviations: AUC = area under the curve, SOTA = state-of-the-art.

Criteria	Precision	Recall	AUC	SOTA AUC
Reached CKD stage 3A	0.951	0.866	**0.965**	0.940
Reached CKD stage 3B	0.900	0.857	0.957	**0.960**
30% eGFR decline	0.846	0.916	**0.952**	0.850

We found that Pearson's r for Mayo predicted eGFR was 0.64. In comparison, our prognosis network's predicted values had an r-value of 0.86 as shown in Fig. 10. Furthermore, the corresponding Bland–Altman plots show that the Mayo imaging classification tool underestimates the absolute predicted eGFR (bias of −1.76 ml/min/1.73 m^2) while our method slightly overestimates the eGFR (bias of 1.18 ml/min/1.73 m^2). Moreover, our approach has a smaller range of 1.96 standard deviations (-26.87 to 29.22 ml/min/1.73 m^2) compared to the Mayo imaging classification tool (-44.88 to 41.37 ml/min/1.73 m^2).

The 8-year eGFR error obtained by our method was not significantly different than that obtained by Mayo classification (two-sided t-test, median difference [interquartile range]: 0.785 [9.95, −5.74] vs 1.85 [12.46, −15.22], p = 0.0548), though the p-value is very close to 0.05 and the difference between the two tools is also clearly shown by Bland Altman plots (Fig. 10).

Finally, our model was also trained to predict each CKD stage distinctly for each patient after 8 years. Here, we reach a weighted f1-score of 0.851 (accuracy: 0.851) with an AUC of 0.972. The corresponding confusion matrix is shown in Fig. 11. As can be seen, most of the predictions are on the diagonal of the confusion matrix. Furthermore, the overall accuracy increases to 0.955 when we factor in misclassified samples from adjacent stages in the confusion matrix.

4 Discussion

In this work, we combined T2-weighted MRIs of the kidneys with the established biomarkers (patient age, eGFR, and predicted HtTKV at the baseline visit) to predict renal function decline. The main advantage of our approach is that our model could classify patients into different CKD stages. This classification might better support the diagnosis of

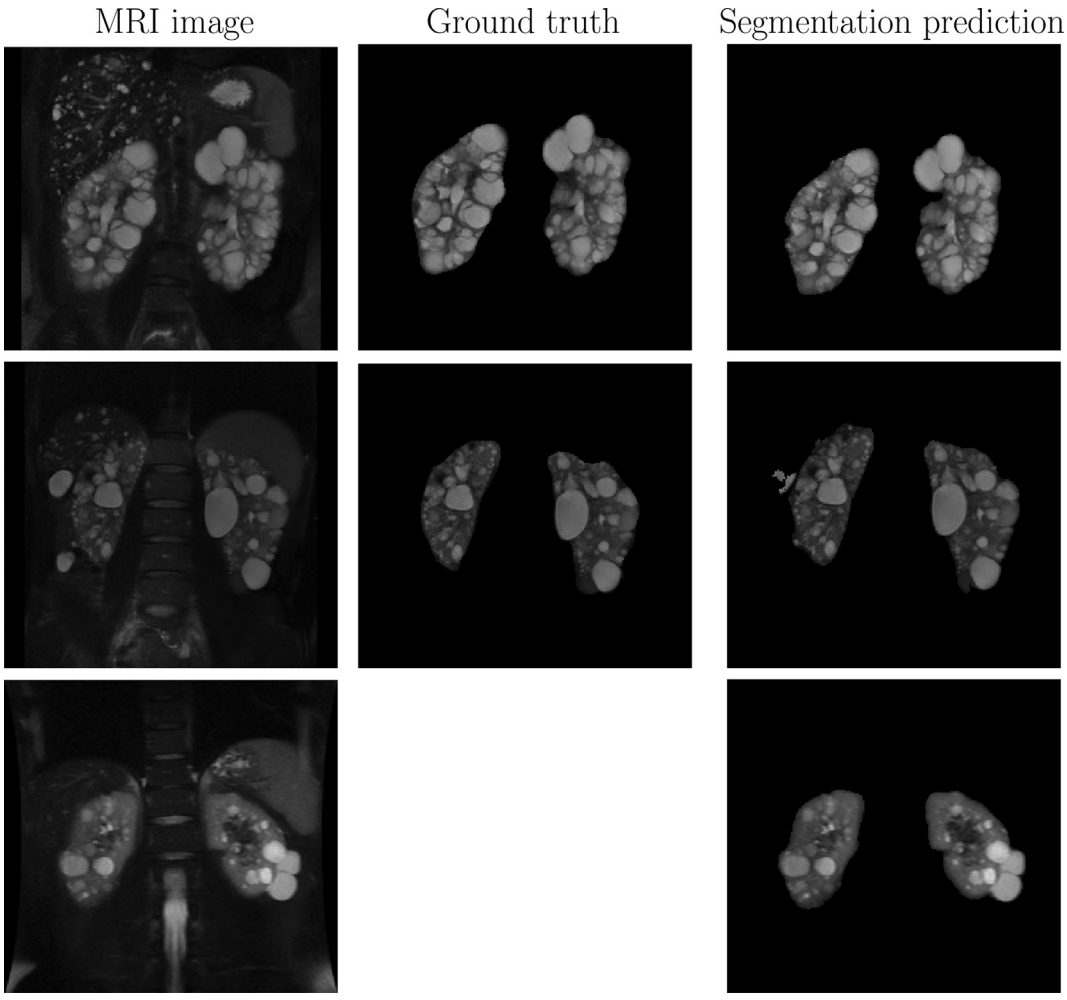

Figure 7. Kidney volume segmentation results. Left: T2-weighted MRI images, center: ground truth segmentations, and right: corresponding automatically segmented kidneys. Three examples show the segmented kidneys obtained from the segmentation network. The first two rows correspond to the cases that had ground truth annotation available, i.e. examples from one of the test set of 100 patients. The last row corresponds to a patient without ground truth segmentation available.

patients with ADPKD since it allows for precisely predicting the change in CKD class and therefore, the decline in renal function over time rather than just predicting if a person will reach a certain CKD stage or not as previously reported in literature [11].

Nevertheless, our model could also demonstrate that it performs at par or better compared to the state-of-the-art method [11] for classification of reaching CKD stage 3A, 3B, and 30% eGFR decline. Furthermore, we also show that an eGFR percent decline could be predicted at a higher rate compared to the Mayo Image Classification tool.

4.1 Kidney segmentation

From a visual inspection, it can be seen that the kidney segmentation is accurate. However, when inspecting the predicted HtTKV vs ground truth HtTKV plot (Fig. 5), we observed that the network predicts slightly higher values, suggesting that it over-segments the kidneys. The mean percent difference of 13.47 $\pm$ 13.70 % between predicted and ground truth HtTKV confirms slight over-segmentation. A reason for this variability could be the slice thickness of 9.0 mm. Nonetheless, we still attained accurate prognosis performance with over-segmented kidneys as inputs to our model. Arguably, over-segmentation is better than under-segmentation since it is more likely to include the kidneys completely. Our segmentation network was originally established on T1 weighted MRIs and there we already showed good performance with respect to previously published studies [15]. In that study, we also noted that some samples contain cysts in various regions of the abdomen and less-defined

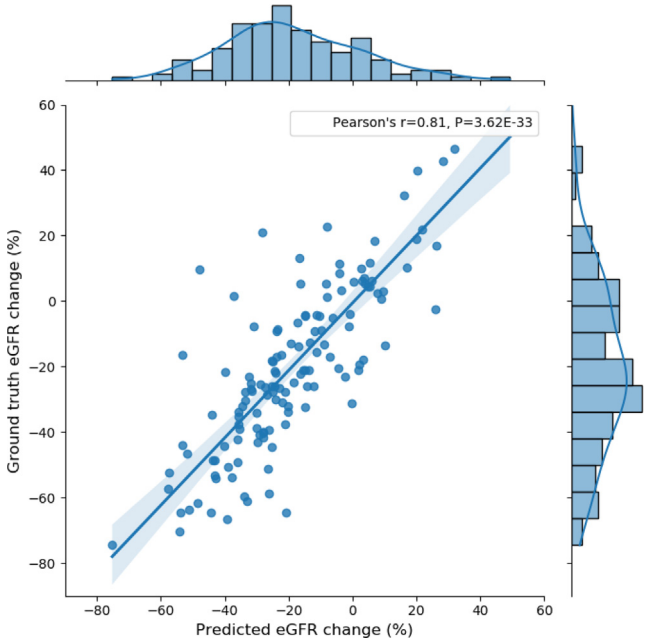

Figure 8. Correlation plot for predicted eGFR percent change v/s ground truth eGFR percent change after 8 years. The Pearson's correlation coefficient is 0.81.

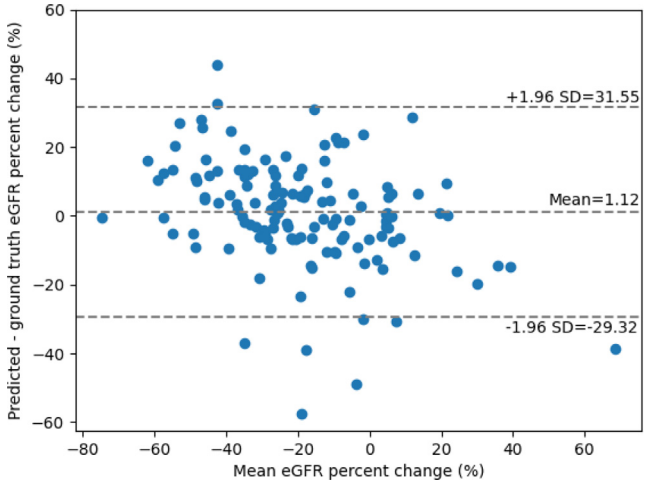

Figure 9. Bland–Altman plot comparing model predicted eGFR percent change (%) values to the ground truth values.

kidney boundaries and shapes that might also lower the segmentation accuracy here. For comparison of T2 weighted segmentation, van Gastel et al. [24] achieved r^2 value of 0.99 between the predicted and ground truth TKV segmentation. In comparison, our approach attains a lower r^2 of 0.96, however, it is worth noting that van Gastel et al. [24] employed four times more data than our method. Furthermore, Mu et al. [25] trained a V-Net [26] like architecture on 305 ADPKD patients. Their study attained a DSC of 0.96, which is higher than our DSC of 0.91. Their DSC score is better than ours since they employ 3 times more data than our study. Nonetheless, the DSC score of our study matches with the inter-user agreement of manual annotation (i.e. 0.91). In another work,Daniel et al. [27] achieved a slightly higher DSC of 0.93 using a similar data size as our study. They employed T2-weighted images with the U-Net. The difference in performance could be due to the subjects in that study consisting of healthy and chronic kidney disease patients only. Those subjects did not have cysts present in the kidney region and hence, segmentation performance would be better in such cases.

4.2 Prognosis

The main objective of our proposed work was to automate ADPKD prognosis and obtain accurate results compared to the work of Kline et al. [11]. Often, we observe that tackling class imbalance is difficult for deep learning models[28–30], however, in this study, we achieved accurate results even though the number of positive samples was considerably less than negative samples, as seen in Tables 1 and 2 (e.g. 23 positive samples v/s 112 negative samples for reaching CKD stage 3B). For the classification tasks (reaching CKD stage 3A, reaching CKD stage 3B, and 30% eGFR decline), our models obtain AUC > 0.950. Our method performs at par in the case of reaching CKD stage 3B classification when compared to the corresponding state-of-the-art results from Kline et al. [11]. Here, we achieve an AUC of 0.957. In comparison, Kline et al. [11] achieve an AUC score of 0.960. It is worth noting that our dataset is slightly more imbalanced than the one used in [11] as depicted in Table 2. Considering this imbalance, our approach is still robust enough in attaining accurate results for every criterion. Note-worthily, our method outperforms the state-of-the-art approach [11] in the cases of reaching CKD stage 3A, 30% eGFR decline, and eGFR percent change. For reaching CKD stage 3A and 30% eGFR decline, our AUCs of 0.965 and 0.952, respectively are higher than that of the state-of-the-art approach [11] (AUCs: 0.940 and 0.850, respectively).

Moreover, our model could be used to classify patients into the five CKD stages (1, 2, 3A, 3B, and 4) after eight years. Our network achieves an AUC and weighted f1-score of 0.972 and 0.851, respectively. We also find that about 95.5% of the predictions lie on the diagonal or the adjacent CKD stages in the confusion matrix (Fig. 11). However, the network misclassifies three cases in the CKD stage 4 category. This occurs due to the class imbalance in the CKD stage 4 category as it consists of only seven cases. Increasing the dataset by including more samples for this class might improve the model's performance.

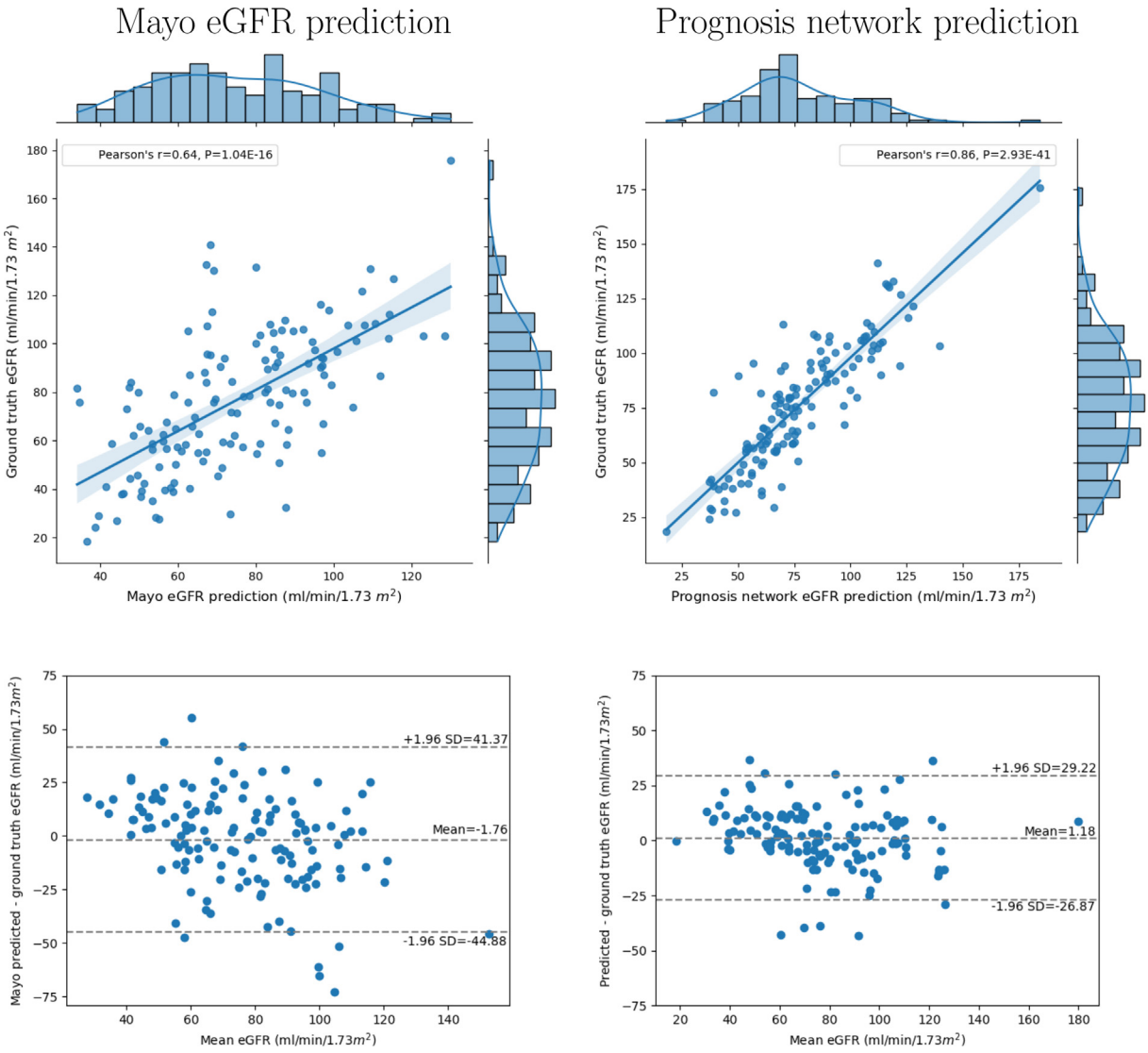

Figure 10. Correlation plots showing predicted versus ground truth eGFR values after 8 years in the first row. The left plot is obtained from Mayo classification tool (Pearsons's r = 0.64) [9,23] and the right plot is from our described prognosis network (Pearson's r = 0.86). The second row depicts the corresponding Bland–Altman plots.

The regression task of eGFR percent change reaches a Pearson's r value of 0.810 as compared to −0.700 of the state-of-the-art method [11]. The features in [11] are negatively correlated to the eGFR percent change, hence, the negative sign. However, our method has a positive correlation since we directly compare the predicted eGFR percent change with the ground truth eGFR percent change (Fig. 8). Furthermore, the residual standard deviation of our approach was also smaller (15.584 % vs 17.900 %). It is clear that the networks are more accurate in predicting eGFR change (classification or regression) after 8 years. Although most of the predicted segmentations have higher HtTKV than ground truth HtTKV, there are 8 cases where there is under-segmentation. This suggests that in this case the predicted segmentations do not cover the kidneys completely. The under-segmentation is on average −20.25 % lesser than the ground truth HtTKV for these cases. Nonetheless, the average absolute difference in eGFR percent change for these eight cases is 9.20%, while the average absolute difference for all 135 patients is 11.50%. This suggests that even though some segmentations do not cover kidneys completely, the other baseline inputs (age, baseline HtTKV, and eGFR) help the prognosis network to predict robust eGFR percent change values. Compared to the Mayo imaging classification tool [9], our algorithm reached a higher correlation of the predicted eGFR to the ground truth

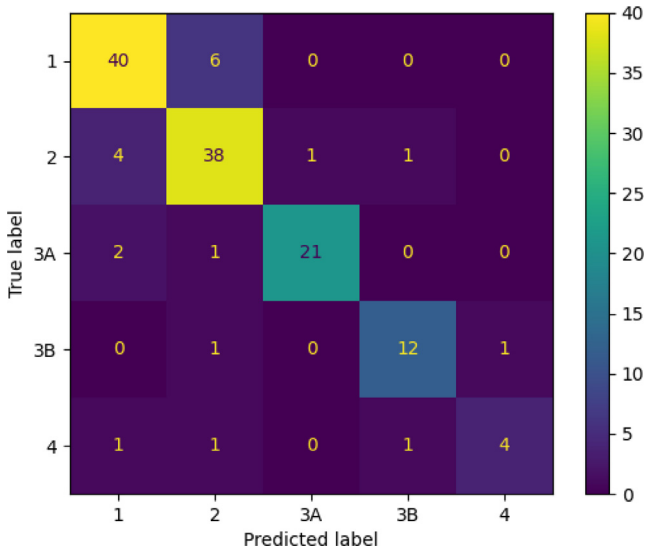

Figure 11. Confusion matrix depicting the predictions for distinct CKD stages after eight years. Weighted f1-score and AUC are 0.85 and 0.97, respectively.

eGFR ($r = 0.64$ vs. $r = 0.86$, Fig. 10). However, the 8-year eGFR error obtained by our method was not significantly different than that obtained by Mayo classification ($p = 0.0548$). Nevertheless, in the light of the correlation and the Bland Altman plot, we believe that including more data might also show significance between the two approaches. Eventually, the main goal was to show the feasibility of automating common approaches to predict disease progression in ADPKD patients. Til now, our tool is on par with the state-of-the-art (Mayo Tool) but more automated.

A few other works exist that predict renal function decline but are not directly comparable to our approach because of differences in datasets and renal function decline definitions. Cao et al. [31] compared various machine learning algorithms on a dataset of 2166 subjects. They defined renal function decline as eGFR decline of more than 3 ml/min/1.73 m^2/year or follow-up eGFR $<$ 60 ml/min/1.73 m^2 (i.e., CKD stage 3A to 5). The algorithms were trained using 24 predictive variables, e.g., age and gender. Their best-performing algorithm was gradient boosting, reaching an AUC of 0.914 on the test data. Zhao et al. [32] explored a prediction model to predict CKD stage 3A or positive proteinuria in a cohort of 348 subjects. Their prediction model combined a genetic risk score model with a non-genetic risk score model by incorporating genetic and non-genetic factors for prediction. This combined model achieved an AUC value of 0.894.

There exist a few approaches that seek to classify distinct CKD stages automatically. However, all approaches do not incorporate imaging data nor do a long-term (larger than 18 months,[33]) prediction of renal function decline. Debal and Sitote [34] created a random forest model that predicts 5 CKD stages (from 1 to 5). They achieved an f1-score of 0.778 with a dataset of 1718 samples. Ilyas et al. [35] employed a dataset of 400 instances and predicted 6 CKD stages (stages 3A and 3B separately). They obtained an overall accuracy of 0.855 using the J48 algorithm, a decision tree based approach. Finally, Rady and Anwar [36] trained a network of probabilistic neural networks on a dataset of 361 patients, achieving an overall accuracy of 0.967. Despite the difference in the underlying data like medical records (such as age, blood pressure, hypertension, hemoglobin, etc.) our integrated approach incorporating imaging information and clinical information (age, eGFR) could achieve a similar or higher classification accuracy while employing only about 3-fold fewer data and allowing for accurate kidney segmentation and TKV estimation simultaneously.

In the future, we plan to use T1-weighted kidney volumes and combine them with their T2-weighted counterparts to extract better features and attempt to improve performance. Moreover, so far we only predicted kidney function change for 8 years in the future. Translating our method to other time ranges will be investigated in the future. Furthermore, we plan to do multi-task learning by combining the training of segmentation and prognosis networks.

In conclusion, we have presented an automated approach to predict disease progression in ADPKD, in terms of eGFR decline and CKD stage change by integrating imaging information and clinical data. Simultaneously, renal segmentations are also obtained and used in further diagnostic tasks. Our approach might improve monitoring and support the prognosis of ADPKD patients from the earliest disease stages.

Data Availability Statement

The code used to extract the data is distributed by the authors as open-source. The patient data cannot be made available on request due to privacy/ethical restrictions.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The Consortium for Radiologic Imaging Studies of Polycystic Kidney Disease (CRISP) was conducted by the CRISP Investigators and supported by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).

The data from the CRISP study reported here were supplied by the NIDDK Central Repository. This manuscript was not prepared in collaboration with investigators of the CRISP study and does not necessarily reflect the opinions or views of the CRISP study, the NIDDK Central Repository, or the NIDDK. We are thankful to the NIDDK for providing us with the patient data from the CRISP study. We gratefully acknowledge the support of the NVIDIA Corporation with the donation of an NVIDIA Titan Xp used for this research. This research project is part of the Research Campus M2OLIE and funded by the German Federal Ministry of Education and Research (BMBF) within the Framework "Forschungscampus: public–private partnership for Innovations" under the funding code 13GW0388A. This study was supported in part by the German Federal Ministry of Education and Research (BMBF) under the funding code 01KU2102, and the Italian Ministry of Health, under the frame of ERA PerMed (ERAPERMED2020-326 - RESPECT). Dr. Caroli acknowledges a grant from the Italian Association for Polycystic Kidney (Associazione Italiana Rene Policistico - AIRP). The authors wish to thank Dr. Norberto Perico, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Bergamo, Italy, for his valuable comments and suggestions.

References

[1] Perrone R, Pirson Y, Schrier R, Torra R, Torres V, Watnick T, Wheeler D. on behalf of the Conference Participants, Autosomal-dominant polycystic kidney disease (adpkd): executive summary from a kidney disease: Improving global outcomes (kdigo) controversies conference. Kidney Int 2015;88:17–27.

[2] Chebib FT, Torres VE. Assessing risk of rapid progression in autosomal dominant polycystic kidney disease and special considerations for disease-modifying therapy. Am J Kidney Dis 2021;78:282–292. https://doi.org/10.1053/j.ajkd.2020.12.020, URL: https://www.sciencedirect.com/science/article/pii/S02726386210043 9X.

[3] Cornec-Le Gall E, Audrézet M-P, Rousseau A, Hourmant M, Renaudineau E, Charasse C, Morin M-P, Moal M-C, Dantal J, Wehbe B, Perrichot R, Frouget T, Vigneau C, Potier J, Jousset P, Guillodo M-P, Siohan P, Terki N, Sawadogo T, Legrand D, Menoyo-Calonge V, Benarbia S, Besnier D, Longuet H, Férec C, Le Meur Y. The propkd score: A new algorithm to predict renal survival in autosomal dominant polycystic kidney disease. J Am Soc Neph 2016;27:942–951. https://doi.org/10.1681/ASN.2015010016, URL: https://jasn.asnjournals.org/content/27/3/942.

[4] Chapman AB et al.. Renal structure in early autosomal-dominant polycystic kidney disease (adpkd): The consortium for radiologic imaging studies of polycystic kidney disease (crisp) cohort. Kidney Intl 2003;64:1035–1045.

[5] Grantham JJ. Polycystic kidney disease: from the bedside to the gene and back. Curr Opin Nephrol Hy 2001;10:533–542.

[6] U. Food, D. Administration, Qualification of biomarker—total kidney volume in studies for treatment of autosomal dominant polycystic kidney disease; 2016.

[7] Cornec-Le Gall E, Alam A, Perrone RD. Autosomal dominant polycystic kidney disease. Lancet 2019;393:919–935. https://doi.org/10.1016/S0140-6736(18)32782-X.

[8] Cornec-Le Gall E, Blais JD, Irazabal MV, Devuyst O, Gansevoort RT, Perrone RD, Chapman AB, Czerwiec FS, Ouyang J, Heyer CM, Senum SR, Le Meur Y, Torres VE, Harris PC. Can we further enrich autosomal dominant polycystic kidney disease clinical trials for rapidly progressive patients? Application of the PROPKD score in the TEMPO trial. Nephrol Dial Transplant 2017;33:645–652. https://doi.org/10.1093/ndt/gfx188.

[9] Irazabal MV, Rangel LJ, Bergstralh EJ, Osborn SL, Harmon AJ, Sundsbak JL, Bae KT, Chapman AB, Grantham JJ, Mrug M, et al. Imaging classification of autosomal dominant polycystic kidney disease: a simple model for selecting patients for clinical trials. J Am Soc Nephrol 2015;26:160–172.

[10] Riyahi S, Dev H, Blumenfeld JD, Rennert H, Yin X, Attari H, Barash I, Chicos I, Bobb W, Donahue S, et al. Hemorrhagic cysts and other mr biomarkers for predicting renal dysfunction progression in autosomal dominant polycystic kidney disease. J Magn Reson Imag 2021;53:564–576.

[11] Kline TL, Korfiatis P, Edwards ME, Bae KT, Yu A, Chapman AB, Mrug M, Grantham JJ, Landsittel D, Bennett WM, et al. Image texture features predict renal function decline in patients with autosomal dominant polycystic kidney disease. Kidney Int 2017;92:1206–1216.

[12] Zöllner FG, Svarstad E, Munthe-Kaas AZ, Schad LR, Lundervold A, Rørvik J. Assessment of kidney volumes from mri: acquisition and segmentation techniques. Am J Roentgenol 2012;199:1060–1069.

[13] Zöllner FG, Koci ski M, Hansen L, Golla A-K, Trbali AŠ, Lundervold A, Materka A, Rogelj P. Kidney segmentation in renal magnetic resonance imaging - current status and prospects. IEEE Access 2021;9:71577–71605. https://doi.org/10.1109/ACCESS.2021.3078430.

[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proc. Intl. Conf. Med. Image Comput. Comput-Assist. Intervent. (MICCAI). Springer; 2015. p. 234–241.

[15] Raj A, Tollens F, Hansen L, Golla A-K, Schad LR, Nörenberg D, Zöllner FG. Deep learning-based total kidney volume segmentation in autosomal dominant polycystic kidney disease using attention, cosine loss, and sharpness aware minimization. Diagnostics 2022;12:1159.

[16] Oktay O, et al. Attention u-net: Learning where to look for the pancreas, arXiv preprint arXiv:1804.03999; 2018.

[17] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. p. 1026–1034.

[18] Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: The missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022; 2016.

[19] Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction, vol. 2. Springer; 2009.

[20] Kingma DP, Ba J. Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980; 2014.

[21] Clevert D-A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus), arXiv preprint arXiv:1511.07289; 2015.

[22] Aggarwal CC. Recommender systems. Springer International Publishing; 2016. https://doi.org/10.1007/978-3-319-29659-3.

[23] Imaging classification of adpkd: A simple model for selecting patients for clinical trials, Accessed 18 October 2022. URL: https://

www.mayo.edu/research/documents/pkd-center-adpkd-classification/doc-20094754.
[24] van Gastel MD, Edwards ME, Torres VE, Erickson BJ, Gansevoort RT, Kline TL. Automatic measurement of kidney and liver volumes from mr images of patients affected by autosomal dominant polycystic kidney disease. J Am Soc Nephrol 2019;30:1514–1522.
[25] Mu G, Ma Y, Han M, Zhan Y, Zhou X, Gao Y. Automatic mr kidney segmentation for autosomal dominant polycystic kidney disease. Proc. Med. Imag. 2019: Comput. Aided Diag, vol. 10950. International Society for Optics and Photonics; 2019. p. 109500X.
[26] Milletari F, Navab N, Ahmadi S-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. 2016 4th Intl Conf. 3D Vis. (3DV). IEEE; 2016. p. 565–571.
[27] Daniel AJ, Buchanan CE, Allcock T, Scerri D, Cox EF, Prestwich BL, Francis ST. Automated renal segmentation in healthy and chronic kidney disease subjects using a convolutional neural network. Magn. Reson. Med. 2021;86:1125–1136.
[28] Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intell Data Anal 2002;6:429–449.
[29] Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019;6:1–54.
[30] Golla A-K, Bauer DF, Schmidt R, Russ T, Nörenberg D, Chung K, Tönnes C, Schad LR, Zöllner FG. Convolutional neural network ensemble segmentation with ratio-based sampling for the arteries and veins in abdominal ct scans. IEEE Trans Biomed Eng 2021;68:1518–1526. https://doi.org/10.1109/TBME.2020.3042640.
[31] Cao X, Lin Y, Yang B, Li Y, Zhou J. Comparison between statistical model and machine learning methods for predicting the risk of renal function decline using routine clinical data in health screening. Risk Manage Healthcare Policy 2022;15:817.
[32] Zhao J, Zhang Y, Qiu J, Zhang X, Wei F, Feng J, Chen C, Zhang K, Feng S, Li W-D. An early prediction model for chronic kidney disease. Sci Rep 2022;12:1–9.
[33] Norouzi J, Yadollahpour A, Mirbagheri S, Mazdeh M, Hosseini S. Predicting renal failure progression in chronic kidney disease using integrated intelligent fuzzy expert system. Comput Math Methods Med 2016:6080814. https://doi.org/10.1155/2016/6080814.
[34] Debal DA, Sitote TM. Chronic kidney disease prediction using machine learning techniques. J Big Data 2022;9:1–19.
[35] Ilyas H, Ali S, Ponum M, Hasan O, Mahmood MT, Iftikhar M, Malik MH. Chronic kidney disease diagnosis using decision tree algorithms. BMC Nephrol 2021;22:1–11.
[36] Rady E-HA, Anwar AS. Prediction of kidney disease stages using data mining algorithms. Informat Med Unlock 2019;15:100178.