Exploring Classification of Histological Disease Biomarkers from Renal Biopsy Images

Puneet Mathur*, Meghna P Ayyar*, Rajiv Ratn Shah
MIDAS Lab, IIIT-Delhi
Delhi, India
pmathur3k6@gmail.com,{meghnaa,rajivratn}@iiitd.ac.in

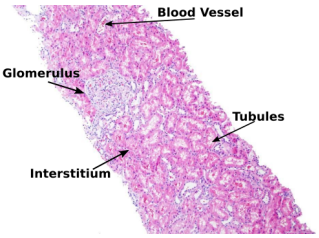
Shree G Sharma
Arkana Laboratories
Arkansas, USA
drshreegopal@gmail.com

Abstract

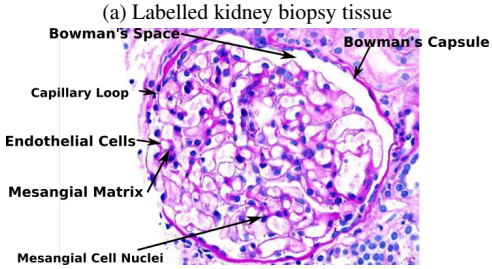
Identification of diseased kidney glomeruli and fibrotic regions remains subjective and time-consuming due to complete dependence on an expert kidney pathologist. In an attempt to automate the classification of glomeruli into normal and abnormal morphology and classification of fibrosis patches into mild, moderate and severe categories, we investigate three deep learning techniques: traditional transfer learning, pre-trained deep neural networks for feature extraction followed by supervised classification, and a novel Multi-Gaze Attention Network (MGANet) that uses multi-headed self-attention through parallel residual skip connections in a CNN architecture. Empirically, while the transfer learning models such as ResNet50, InceptionResNetV2,VGG19 and InceptionV3 acutely under-perform in the classification tasks, the Logistic Regression model augmented with features extracted from the InceptionResNetV2 shows promising results. Additionally, the experiments effectively ascertain that the proposed MGANet architecture outperforms both the former baseline techniques to establish the state of the art accuracy of 87.25% and 81.47% for glomeruli and fibrosis classification, respectively on the Renal Glomeruli Fibrosis Histopathological (RGFH) database.

1. Introduction

A kidney tissue comprises of multiple functioning units called nephrons, which are comprised of glomeruli and tubules. The area in-between the tubules is known as the interstitium. Glomeruli are the principal filtering units of a kidney and most of the renal diseases affect the glomerular segments [14]. Glomeruli exhibit high variability in terms of size, shape, and color, even in the same tissue sample. This is fundamentally due to their relative position and alignment, heterogeneity in staining and genetic biological processes. Generally, glomeruli are spherical in shape and may be distorted in disease conditions, e.g., hypertension and diabetes. Any change in the shape, cellularity, size or structure of the glomeruli might act as one of the early indicators of kidney diseases.



(a) Labelled kidney biopsy tissue



(b) Labelled glomerulus

Figure 1: Features of kidney tissues identified by a nephropathologist

A glomerulus is marked as abnormal if there is a deviation from normal morphology and staining characteristics [30]. Figure 1 shows the basic parts of a kidney tissue and an annotated glomerulus. A normal and healthy glomerulus shows no expansion of mesangial matrix and cellularity. The glomerular capillary loops are patent and the glomerular basement membranes appear to be of normal thickness. There is no proliferation in the Bowman space[1], no necrosis is seen and glomerular tuft is not ad-

*Authors contributed equally

[1]Sack like structure in the kidneys that performs blood filtration.

Pre-print of paper accepted at WACV 2019

hered to the Bowman capsule. An abnormal glomerulus represents a departure from normal histology in terms of sclerosis (stiffening of the glomerulus caused due to the replacement of the original tissues by connective tissue), a proliferation of glomerular capillaries and endothelial cells, infiltrating leukocytes and obliteration of capillary spaces.

Diseased glomeruli form a scar upon healing, called fibrosis, similar to a wound healing on the skin after a cut followed by replacement of normal skin. In kidneys, fibrosis replaces the functioning nephrons and these scarred tissues do not contribute to the functioning of a nephron. Therefore, nephrons once damaged cannot be replicated, making the scar tissue or the damage to the kidney irreversible. The frequency of diseased glomeruli and extent of renal fibrosis act as hallmarks of the underlying progression of chronic kidney disease (CKD), and the medical prognosis. Diagnosing kidney health from kidney biopsies is very subjective and requires the presence of an expert to provide a proper diagnosis. Therefore, we aim to automate the diagnosis on kidney biopsy by using different deep learning based techniques, so as to alleviate this dependence on the presence of an expert and also make the process more objective. As a preliminary step in this direction, we accomplish the classification of renal glomeruli into two fundamental classes: normal and abnormal. Detecting the presence of abnormal glomeruli in the renal tissue slide is the most basic step that a pathologist performs to decide whether the tissue is affected or healthy. Alongside, the work explores the identification of patches of fibrotic renal biopsy into three elementary classes: mild, moderate and severe to determine the progression of renal disease.

Transfer learning models, ResNet50, InceptionV3, InceptionResNetV2 and VGG19 have been engaged in our work as baselines similar to the earlier work in glomureli classification [3]. Secondary supervised classifiers used in this study are Logistic Regression (LOGREG) [2], Random Forest (RF) [6] and Naive Bayes (NB) [40]. Each of these classifiers takes the feature vectors of the image data as the input that is extracted from the deepest layers of the respective pre-trained image classification models- ResNet50, InceptionV3, InceptionResNetV2, and VGG19 by convoluting the native image RGB descriptors with the weights of the last layers of individual architectures. The respective pre-trained architectures, when used as feature extractors, are referred to as IRFE (InceptionResNetV2 Feature Extractor), IFE (InceptionV3 Feature Extractor), RFE (ResNet50 Feature Extractor) and VFE (VGG19 Feature Extractor) throughout the paper. Lastly, we introduce a self attention based neural architecture, known as Multi-Gaze Attention Network (MGANet). The main contributions of this study can be summarized as follows:

- Creation of Renal Glomeruli Fibrosis Histopathology database (RGFH) consisting of two datasets: Renal Glomeruli Dataset (RGD) and Renal Fibrosis Dataset (RFD). The dataset images have been collected from whole slide images(WSIs) or static images taken by multiple in-house experts and then verified by our expert kidney pathologist.

- Experimentation to ascertain the applicability of simple transfer learning using ResNet50, InceptionV3, InceptionResNetV2 and VGG19 models.

- Experimentation to analyze the performance of supervised secondary classifiers including Logistic Regression, Random Forest and Naive Bayes that use weighted image feature vectors from pre-trained transfer learning architectures such as ResNet50, InceptionV3, InceptionResNetV2 and VGG19 architectures as inputs.

- Investigation of the proposed MGANet for classification of glomeruli and fibrotic images. We incorporate scaled dot product attention and draw a comparison of the relative arrangement of input attention maps for optimal performance. We also try to figure out the most promising deep neural network that provides optimal performance on classification tasks.

The rest of the paper is organized as follows. The important related work is reported in Section 2 followed by detailed discussion of RGFH database in Section 3. Section 4 introduces the proposed methodology. The experimental results and comparison with the state-of-the-art methods are mentioned in Section 5 and comprehensive error analysis is reported in Section 6. Finally, Section 7 concludes and suggests future work.

2. Related Work

Simple supervised classification techniques involving SVM [19] and Gradient Boosting Decision Tree [27] have been successful on textual data but not so much in the domain of image classification. On the other hand, deep learning paradigms take advantage of the massive amount of training data in conjugation with their inherent neural architecture to investigate the data complexities without an auxiliary understanding of the nuances of the medical field. Shin *et al.* [32] gave a descriptive explanation of the applications of transfer learning from pre-trained ImageNet [10] based frameworks to an allied image corpus. A detailed mathematical analysis of feature extractors was put forth by [37], that inspected the idea of feeding characteristic features of the signals to improve classification performance. ResFeats put forth by Mahmood *et al.* [17] portrayed the usefulness of pre-trained ResNet based feature extractor over multiple datasets as a remarkable improvement in object classification, scene classification and coral classifica-

Pre-print of paper accepted at WACV 2019

tion tasks. Deep cascaded networks were employed on routine HE stained tissues to detect mitosis in breast cancer tissues Chen *et al.* [8]. Locality sensitive deep neural network frameworks have also been utilized for automatically detecting and classifying individual nuclei in colon histology images [33]. Convergent approaches have been tried earlier to combine domain inspired features with CNN's to detect mitosis, thereby reducing the excessive dependency on large datasets and associated intuition on deep learning frameworks [36]. Regions of prostate cancer were then classified via boosted Bayesian multi-resolution classifier followed by applying Gabor filter features using an AdaBoost ensemble method [11].

3. Renal Glomeruli-Fibrosis Histopathological (RGFH) Database

The RGFH database comprises two datasets: Renal Glomeruli Dataset (RGD) and Renal Fibrosis Dataset (RFD). RGD consists of glomeruli images partitioned into normal and abnormal classes. RFD dataset consists of kidney tissue images partitioned into mild fibrosis, moderate fibrosis and severe fibrosis classes. The constituent de-identified images of both the datasets have been sourced from Arkana Laboratories[2] after seeking prior approval from the ethics committee to avoid privacy concerns and following patient anonymity rules.

3.1. Database Acquisition

The de-identified biopsy images, similar to Figure 1 were procured between January 2018 to July 2018. The kidney tissues have been extracted through needle biopsies and were processed and stained according to published standards [9]. The tissue samples were digitized using MoticEasyScan[3] at 20X (0.5 micron/pixel). The images are obtained in TIFF-based SVS format that was converted into JPEG format. The scanner was equipped with 15 fps 2/3" CCD sensor and comes fitted with CCIS Infinity optics for reliable, fast and efficient work in cytology, histology and cytopathology. The static images of glomeruli and tissue patches were captured using the Olympus camera. The digitized biopsy images consist of an amalgamation of several renal substructures such as interstitial tissue, tubules, blood vessels and glomeruli [15] at 20X and 40X. Images having insufficient staining, poor light intensity and fragmented tissue portions were not included in the dataset.

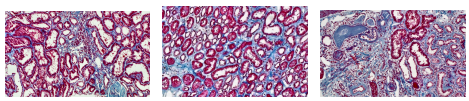
3.2. Database Preparation

RGD consists of independent sections of segmented glomeruli taken from static images of kidney biopsies at a uniform 40x magnification. The procured patches of glomeruli were subjected to further filtering where glomeruli images with missing borders, insufficient staining, poor light intensity and fragmented glomeruli portions were removed. The content and quality of the remaining images were verified to have adequate pixel intensity, contrast and minimal blurring.

The presence of heterogeneous substructures and multiple tissue constructs in a particular WSI compound the task of identifying the fibrotic region in the tissues for the RFD dataset. As a result, assigning the fibrosis label becomes a complex process. The whole slide images were broken down into an array of a large number of rectangular windows, each referred to as a *patch*. While extracting the renal tissue patches, each of the patches was so chosen to have less than 10% non-tissue region.

(a) Abnormal glomerulus (b) Normal glomerulus

Figure 2: Examples of RGD for each class



(a) Mild (b) Moderate (c) Severe

Figure 3: Examples from RFD for each class

3.3. Database Statistics

RGD is a dataset of 935 images of renal glomeruli obtained from human renal biopsies. The dataset has been constructed to have 619 abnormal and 316 normal images for the respective class labels (see Figure 2). An important characteristic feature of the dataset provided is the use of multiple stains that resembles standard clinical practice along with the presence of subtle non-uniformity in the degree of staining which emphasizes the natural unavoidable variations in biopsy processing throughout the medical world.

RFD dataset follows an annotation schema wherein the entire dataset of 927 images is formulated into three classes based on the extent of regional scarring: (i) mild ($5-25\%$), (ii) moderate ($26-50\%$) and (iii) severe (more than 50%). The dataset contains 356 samples of mild fibrosis, 198 samples of moderate fibrosis and 373 samples of severe fibrosis (see Figure 3). Tables 1 and 2 describe the distribution of the classes in each dataset.

[2] www.arkanalabs.com
[3] https://www.motic.com

Pre-print of paper accepted at WACV 2019

Label	Count
Abnormal	619
Normal	316
Total	935

Table 1: Image label distribution in RGD

Label	Count
Mild	356
Moderate	198
Severe	373
Total	927

Table 2: Image label distribution in RFD

3.4. Database Annotation Protocol

As per the sourcing laboratory, all the tissue samples and their annotations meet the medical standards set by responsible accreditation bodies and were cross-annotated by multiple pathologists during the real-time patient-testing phase to ascertain their credibility in clinical diagnosis. The RGFH database was extracted from a pool of archived renal WSIs or static images data available with the laboratory. The expert kidney pathologist responsible for the verification of the image annotations, having an extensive background in kidney pathology, verified those images and their corresponding annotations made by multiple in-house pathologists on real-world medical cases. In this way, the images were exposed to another round of scrutiny and factual validation, diminishing any chance of incorrect annotations. At each step, the dataset was subjected to due medical diligence [35] in consensus with the kidney pathologist.

4. Methodology

The following section is divided into four parts: Section 4.1 describes the pre-processing steps applied to images in the RGFH database, followed by Section 4.2 and 4.3 which highlight the transfer learning and supervised classification with DNN feature extraction respectively. Finally, Section 4.4 covers the discussion on the proposed Multi-Gaze Attention Network(MGANet) model.

4.1. Data Pre-processing

To keep the model invariant to fine changes in image quality, we perform certain pre-processing steps like histogram equalization for enhancement of image contrast. Contrast-Limited Adaptive Histogram Equalizer (CLAHE) [25] was used for pre-processing both RGD and RFD dataset. Realizing the problems associated with the small size of the proposed database, data augmentation techniques were applied to handle data inadequacy, data imbalance and lack of uniform modalities across the datasets. Along with data warping and synthetic oversampling [38], elastic deformations were also employed to generate plausible transformations of existing samples without distorting the original label information. Table 3 lists all the techniques used.

Parameter	Value
Horizontal Flip	True
Vertical Flip	True
Fill Mode	Nearest
Zoom Range	0.1
Width Shift Range	0.2
Height Shift Range	0.2
Rotational Range	180

Table 3: Data augmentation parameters

4.2. Transfer Learning

Training a CNN directly from scratch requires significantly greater time and training data [21]. Alternatively, fine-tuning the pre-trained models in case of similar base and target data remarkably enhances the generalization performance of the classifier [39]. Esteva et al. [12] demonstrated the classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs to InceptionV3. Inspired by the same, we explored several transfer learning architectures such as ResNet50, InceptionV3, InceptionResnet and VGG19 models initialized with corresponding ImageNet weights. As depicted in Figure 4, the models are re-trained by freezing the weights of all trainable layers except the last three dense fully-connected layers. The activation function applied is ReLU [16] for the second and third last dense layers followed by 'Softmax'in the last dense layer.

Let domain D consist of two components: a feature space X and a marginal probability distribution $P(X)$. x_i represents the input image and y_i represents the output label corresponding to the sample image from the RGFH dataset. Z represents the pre-trained weights of ImageNet classification. Transfer learning framework T, mathematically outputs a predicted label space through the transfer function f, which is retrained on the data pairs of (x_i, y_i) as shown in Equation (1). It takes in the tuples of the image and label along with pre-initialized layer weights and the output vector of class probabilities in the form of Y is shown in Equation (2), where each class label is distinctly referred to as $\alpha, \beta \ldots$

$$Timage = f(Z, x_i, y_i) \tag{1}$$

$$Y_i = \{{P_i}^{\alpha}(T(x_i)), {P_i}^{\beta}(T(x_i)) \ldots\} \tag{2}$$

The batch-size and epochs were chosen by grid search in the range of (8-128) and (20-100) in equal spaced intervals through 5-fold cross-validation for optimal performance. The final models had a batch size of 16 and were trained for 30 epochs. The loss function was chosen as categorical cross-entropy with the Adam optimizer and L2 normalization.

Pre-print of paper accepted at WACV 2019

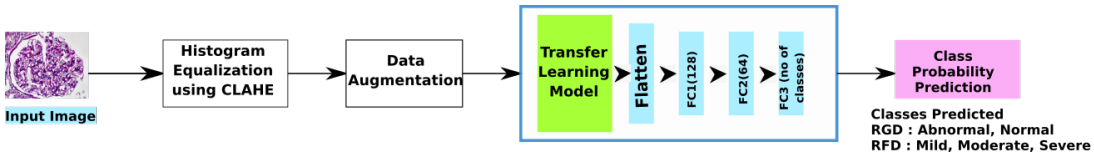


Figure 4: Framework of transfer learning model [3]

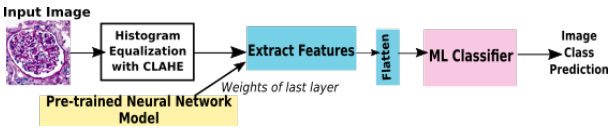


Figure 5: Framework of supervised classification with feature extraction [3]

4.3. Supervised Classification with DNN Feature Extraction

The features extracted at the deeper layers of the CNN usually correspond to the subtle intricacies and are characteristic of medical image datasets. These features are predominantly difficult to detect through regular image descriptors as the images show minute variations across the spectrum of the dataset. On the other hand, the first few layers contain generic features resembling Gabor filters or blob features [39].

The activations of the fully connected (FC) layers capture the overall tissue level substructures, while the last output layer preserves the spatial information representing particular classes in datasets. The loss of local spatial information such as the several substructures of the biopsy tissue during propagation through the fully connected layers, which explains the reasoning behind reusing a pre-trained DNN as a feature extractor, instead of using it directly as a classifier [22].

Also, features obtained through last layers of pre-trained CNN outperform SIFT and HOG image descriptors [24]. Thus, there exists a possibility to investigate similar methodologies to exploit supervised classification by Logistic Regression, Naive Bayes and Random Forest respectively on pre-trained networks adopted for the feature extraction in our work including InceptionV3, Inception-ResNet, ResNet50 and VGG19 in manner similar to that illustrated in Figure 5. The proposed strategy starts with extraction of last level feature vectors from a pre-trained CNN model followed by matrix multiplying the obtained weights with image vectors to form image specific feature vectors. The output of the feature extractor is generally of the form $w * h * d$ where w is the width, h is the height and d is the number of channels in the convolutional layers. These 2D arrays of d dimension are flattened and trained on Logistic Regression, Naive Bayes and Random Forest models with the image feature vectors as an input.

The complete proposed strategy can be understood by the following mathematical description. Let us assume X as the sample space of all input training data from RGFH. The input-output pairs of image and the corresponding label are represented as (x_i, y_i). $P(X)$ is the probability function of the output labels. As per Equation (3), the image vector of a sample image $\rho(x_i)$ is convoluted (λ) with the last layer weights of the pre-trained model $\tilde{Z}$ to give the transformed input (ν) for the secondary classifier. The secondary supervised classifier g is trained with modified input image $\nu(x_i)$ and the corresponding label y_i by passing through the supervised classifier as shown in Equation (4). Equation (5) demonstrates that the output label space Y is a vector of probabilities of distinct label classes $(\alpha, \beta \ldots)$ obtained when the trained classifier C is tested on the image samples.

$$\nu(x_i) = \lambda(\tilde{Z}, \rho(x_i)) \tag{3}$$

$$C_{image} = g(\nu(x_i), y_i)) \tag{4}$$

$$Y_i = \{P_i^{\alpha}(C(x_i)), P_i^{\beta}(C(x_i)) \ldots\} \tag{5}$$

The final layers $\tilde{Z}$ of pre-trained models from which the feature weights were extracted are given in Table 4. The hyper-parameters for random forest classifier were fine-tuned using 10-fold cross-validation and the results were found to be optimal when n_estimators, max_depth and max_features were fixed at 1000, 15 and $log2$ respectively.

Pre-trained architecture	Layer
IRFE	CONV_7B
IFE	MIXED10
RFE	AVG_POOL
VFE	FC1

Table 4: Layers contributing to image feature weights.

4.4. Multi-Gaze Attention Network (MGANet)

In bio-medical image classification, tissue level discriminating features are generally localized rather than being present in the entirety of the image. Pre-trained deep learning models are unable to prioritize features extracted from relevant local patches, thereby focusing on global pixel level information. Figure 1 highlights the intuitive handcrafted features used by clinicians to derive disease prognosis. The problem can be effectively handled by extracting handcrafted features from corresponding biopsy images but the subjectivity involved in process outweighs the benefit

Pre-print of paper accepted at WACV 2019

of feature interpretability. Alternatively, an unsupervised feature generation approach of attention based deep learning strategies [23] can facilitate seamless domain adaption, irrespective of the fundamental disease characterization of the image.

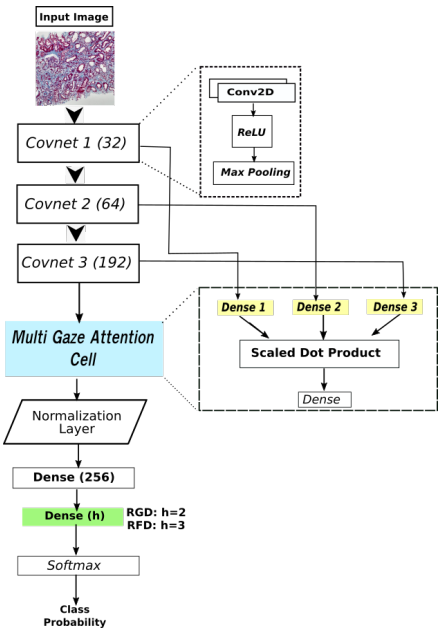


Figure 6: Framework of MGANet. (32, 64 and 192 are the sizes of the Conv2D of respective Covnets)

Figure 6 portrays the proposed MGANet consisting of three ConvNets, each composed of a Convolutional 2D layer with ReLu activation followed by a maxpooling layer having a stride of 2 units. Further, three corresponding attention maps are generated over the whole slide image through residual skip connections from the initial ConvNet layers in parallel. Each of the three attention maps is appended and their scaled dot product is computed. The scaled dot product of feature maps is preferred over additive and simple attention due to its space-efficiency [7]. It allows the model to jointly attend to information from different representation sub-spaces at different positions. Each attention map is treated as an $l * l$ matrix of blocks such that each block is of dimension $d * c$, where d represents the output dimension of the last convolutional layer and c represents the number of input color channels (RGB, *i.e.,* 3 in our case). Each block of attention map is passed through a fully connected dense layer and the output obtained is reshaped to dimension $d_v * n_v$, where d_v and n_v represent the dimension of the linear space the input is to be projected and the number of projections for each block respectively. This gives rise to multi-headed self-attention within the attention maps inspired by [34]. A triplet of the three distinct attention maps, a_1, a_2 and a_3, is then passed through the scaled dot product function where the scaled product of vectors a_1 and a_3 is passed to get the Softmax over vector a_2, as shown in Equation (6). The permutation of the attention maps $\{a_1, a_2, a_3\}$ is manipulated to derive the best possible experimental configuration. The resulting output is layer normalized [4] to make the tensor have a standard normal distribution by deleting some dimensions of the vector that are not important. This can be viewed as another smaller attention in itself at the normalization stage. The final output is then flattened and passed through two consecutive dense fully connected layers with activation functions ReLU and Softmax respectively.

$$Attention(a_1, a_2, a_3) = \frac{Softmax(a_1, a_3) * a_2}{\sqrt{max(a_1, a_3)}} \quad (6)$$

5. Results and Discussions

Dataset		Resnet50	IV3	IRV2	VGG19
RGD	*Acc*	**72.35**	68.56	64.75	69.89
	Prec	74.56	71.35	60.12	66.59
	Recall	72.56	68.92	65.81	68.71
	F1	72.97	69.01	55.41	64.90
RFD	*Acc*	**66.87**	52.47	53.64	64.74
	Prec	51.40	41.38	42.69	53.15
	Recall	62.12	50.67	51.21	61.72
	F1	60.25	48.76	42.81	59.83

Table 5: Results (in %) for transfer learning model. (IV3: InceptionV3, IRV2: InceptionResnetV2, Acc: Accuracy, Prec: Precision)

Table 5 reports the results of transfer learning methodology as discussed in Section 4.2. Experimental results were obtained by using weighted metrics in each case as the class imbalance may skew the model performance to unilaterally favor the dominant class. The models in each case were trained with stratified K-fold cross-validation with K=5, for parameter tuning to further account for class imbalance. The preliminary results gathered from transfer learning models aim to serve as a baseline for classification for features extraction from pre-trained DNN and MGANet. The results derived from both the datasets seems to support our hypothesis that transfer learning models unexpectedly suffer from misclassification due to high congruity and intricate variations in biopsy image. Amongst ResNet50, InceptionResNetV2,InceptionV3 and VGG19, ResNet50 clearly outperforms contemporary pre-trained models significantly due to the presence of shortcut connections known as residual networks in its architecture. ResNet50 architecture records the best accuracy of 72.35% and 66.87% on RGD and RFD datasets respectively and a similar trend is prevalent across other metric measurements too.

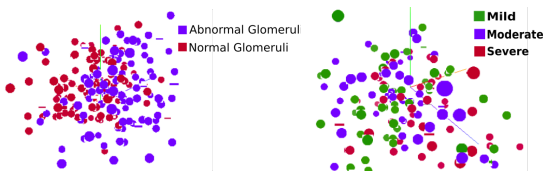
Pre-print of paper accepted at WACV 2019

Classifiers		Acc	Prec	Recall	F1
IRFE	*LogReg*	**85.23**	**85.64**	**86.03**	**85.08**
	Random Forest	80.74	83.08	76.74	78.11
	Naive Bayes	72.19	74.80	72.89	72.94
IFE	*LogReg*	81.63	82.53	78.63	80.35
	Random Forest	80.21	84.76	74.21	77.38
	Naive Bayes	75.42	75.40	75.40	75.40
RFE	*LogReg*	83.42	83.16	83.42	83.07
	Random Forest	82.49	82.74	84.49	83.41
	Naive Bayes	67.91	71.39	67.91	68.73
VFE	*LogReg*	83.16	83.02	84.16	83.99
	Random Forest	83.70	82.62	81.70	82.14
	Naive Bayes	60.42	74.39	60.42	60.99

Table 6: Results (in %) of supervised classification with pre-trained feature extractor for RGD (LogReg: Logistic Regression, Acc: Accuracy, Prec: Precision).

Classifiers		Acc	Prec	Recall	F1
IRFE	*LogReg*	**71.51**	59.74	**68.51**	**69.96**
	Random Forest	66.45	43.95	56.45	59.34
	Naive Bayes	54.62	44.97	44.62	54.71
IFE	*LogReg*	64.81	60.89	64.81	64.38
	Random Forest	59.67	56.61	59.62	56.82
	Naive Bayes	61.82	50.09	61.81	55.15
RFE	*LogReg*	58.06	51.33	58.06	53.89
	Random Forest	58.06	61.31	58.06	52.62
	Naive Bayes	53.09	43.64	37.09	38.82
VFE	*LogReg*	69.74	**66.89**	67.74	67.10
	Random Forest	67.21	56.01	60.21	55.31
	Naive Bayes	60.53	53.53	50.53	51.44

Table 7: Results (in %) of supervised classification with pre-trained feature extractor for RFD (LogReg: Logistic Regression, Acc: Accuracy, Prec: Precision).



(a) (b)

Figure 7: t-SNE plots for (a) RGD and (b) RFD

Feature extraction based supervised classification was tested on both the datasets and the corresponding results were compiled in Tables 6 and 7. The findings give plausible credibility to our proposition to use a CNN-based feature extraction as a preliminary step for the WSI classification of RGD and RFD datasets. Out of multiple pre-trained models used on different classifiers, the Logistic Regression supplemented with features imparted by InceptionResNetV2 is most successful in terms of accuracy, recall and F1 score. A general inference is that the feature extraction based classifiers showed an improvement in contrast to transfer learning based methods. Among contemporary secondary classifiers, Logistic Regression shows a greater ability to adapt to the fine-grained features of the glomeruli images with high capacity to fit onto a small-sized, immensely correlated dataset. On the other hand, the InceptionResNetV2 framework gives the most favorable results compared to its corollaries followed by VGG19 as evident by slight degradation in the metric values.

Table 8 details the performance of attention based MGANet on RGD and RFD for various permutations of scaled dot product. It can be directly observed that the model outperforms both transfer learning as well as feature extraction based supervised classification techniques by a significant margin, substantiating our claim that localized features present in medical histological images are more prominent than universal pixel-level image features. Interestingly, changing the relative order of attention maps used for calculating the scaled dot product attention did not report a convincing deviation in performance metrics. The t-SNE plots of RGD and RFD datasets, as shown in Figure 7, support the argument that the inter-class heterogeneity amongst class labels is low. The high overlap of clusters point to the fact that the tissue level micro-structural differences amongst the constituent classes are subtle and require intricate feature modeling for the classifiers to take them into consideration. This justifies the claim and the related observation that classification on image features extracted through pre-trained transfer learning models perform better than naive transfer learning. Further, the facilitation of unsupervised localized attention through MGANet serves a role similar to using handcrafted features. Thus, it can be summed up that feature extraction using InceptionResNetV2 model performs relatively better than baseline transfer learning with ResNet50, while the MGANet surpasses both the methods on both RGD and RFD.

Classifier	Accuracy	Precision	Recall	F1
RGD $(\alpha_1, \alpha_2, \alpha_3)$	**87.25**	**75.91**	**87.99**	**87.17**
RGD $(\alpha_2, \alpha_3, \alpha_1)$	87.17	75.89	87.56	87.14
RGD $(\alpha_3, \alpha_1, \alpha_2)$	87.08	75.88	87.49	87.09
RFD $(\alpha_1, \alpha_2, \alpha_3)$	**81.47**	**58.77**	**84.23**	82.64
RFD $(\alpha_2, \alpha_3, \alpha_1)$	81.41	58.56	84.11	81.64
RFD $(\alpha_3, \alpha_1, \alpha_2)$	81.38	58.48	84.15	**82.79**

Table 8: Results (in %) for MGANet

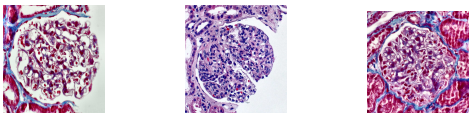
6. Error Analysis

A brief analysis is presented in this section highlighting various limitations encountered while classifying RGD and RFD images along with suggested improvements.

1. **High variability in staining:** In routine clinical prac-

Pre-print of paper accepted at WACV 2019

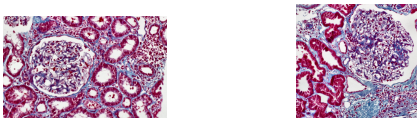
tice, a kidney biopsy after processing is stained with Hematoxylin and Eosin (HE), Periodic acid-Schiff (PAS), Jones Methenamine Silver (JMS) and Masson trichrome (MT) stain [13]. To render a comprehensive clinical diagnosis all the stains are used in conjunction and many more subtle features are interpreted besides the ones mentioned above. Although each pathology lab follows standardized practices in terms of chemical composition and procedure for tissue staining, still there is a lot of variability in staining from lab to lab and case to case basis. Figure 8 shows samples of whole slide images and their corresponding ground truth annotations misclassified by ResNet50, InceptionResNetV2 based feature extraction and MGANet respectively due to the same.



(a) Normal (b) Abnormal (c) Abnormal

Figure 8: Illustration of RGD images with true labels misclassified by (a)Resnet50 (b) InceptionResNetV2 based feature extraction (c) MGANet

2. **Low precision in RFD dataset image classification:** Although the ranking order of proposed methods is consistent in case of both RGD and RFD, a prominent exception of considerably low precision exists in case of RFD. The primary cause for the same is the problem of the presence of heterogeneous tissue-level noise. This can be attributed to the fact that the biopsy images constituting the RFD dataset are not devoid of glomeruli segments, leading to erroneous classification as evident from Figure 9 which exhibits images consistently misclassified by the models. Currently, our work does not deal with automatic segmentation of glomeruli and will be addressed in the future.



(a) Mild Fibrosis (b) Moderate Fibrosis

Figure 9: Given true labels, misclassified as severe

7. Conclusion and Future Work

Kidney health bio-markers can aid in pre-assessing the progression of potentially fatal chronic kidney diseases, better quantitative characterization of disease and precision medicine. Challenges such as the absence of a reliable biomedical dataset of glomeruli images, challenges in biopsy digitization, heterogeneity across entire patches of the tissue section, color variations induced due to differences in slide preparation and the existence of a deluge of deep learning options for the image classification tasks require concentrated efforts in dataset acquisition, preparation and investigation of a computationally inexpensive technique while ensuring medical trustworthiness at the same time. Most medical datasets suffer from a peculiar gold standard paradox [1]. The experiments prove that vanilla transfer learning models fail to surpass feature-enriched linear classification models owing to high interclass similarities. Development of suitable fine-tuned algorithms that do not converge to the set biases posed a challenging task that eventually abated the usage of widely popular transfer learning and pre-trained image feature extractors on highly subjective medical datasets. In order to outperform the baseline models, Multi-Gaze Attention Model (MGANet) was introduced to replace the cumbersome feature extraction with unsupervised multi-headed self-attention followed by scaled dot product.

The current findings aim to establish a state of the art in the novel area of renal histopathology. The RGD dataset can be extended to include unreported categories of glomeruli such as Sclerotic and Crescentic [5]. A potential advancement in precision metrics for classification of fibrosis images can be the use of stacked convolutional auto-encoders for hierarchical feature extraction as depicted in similar domains [18]. Moreover, advanced neural architectures such as bi-channel CNN-LSTM models [20] and C-LSTM's [26]. Wrapper-penalty based feature selection algorithms can also be utilized for choosing the best possible set of features suitable for efficient classification [28, 29]. A potential advancement can be the extension of this work by incorporating clinical parameters of users through multimodal devices [31].

Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation for the donation of a Titan Xp GPU used for this research.

References

[1] F. Aeffner, K. Wilson, N. T. Martin, J. C. Black, C. L. L. Hendriks, B. Bolon, D. G. Rudmann, R. Gianani, S. R. Koegler, J. Krueger, et al. The gold standard paradox in digital image analysis: manual versus automated scoring as ground truth. *Archives of pathology & laboratory medicine*, 141(9):1267–1275, 2017.

[2] A. Agresti. *Logistic regression*. Wiley Online Library, 2002.

[3] M. P. Ayyar, P. Mathur, R. R. Shah, and S. G. Sharma. Harnessing ai for kidney glomeruli classification. In *Proceedings of 20th IEEE International Symposium on Multimedia*, 2018.

Pre-print of paper accepted at WACV 2019

[4] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[5] A. E. Berden, F. Ferrario, E. C. Hagen, D. R. Jayne, J. C. Jennette, K. Joh, I. Neumann, L.-H. Noël, C. D. Pusey, R. Waldherr, et al. Histopathologic classification of anca-associated glomerulonephritis. *Journal of the American Society of Nephrology*, 21(10):1628–1636, 2010.

[6] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *IEEE 11th International Conference on,Computer Vision (ICCV)*, pages 1–8. IEEE, 2007.

[7] D. Britz, A. Goldie, M.-T. Luong, and Q. Le. Massive exploration of neural machine translation architectures. *arXiv preprint arXiv:1703.03906*, 2017.

[8] H. Chen, Q. Dou, X. Wang, J. Qin, P.-A. Heng, et al. Mitosis detection in breast cancer histology images via deep cascaded networks. In *AAAI*, pages 1160–1166, 2016.

[9] J. Churg and M. A. Gerber. The processing and examination of renal biopsies. *Laboratory Medicine*, 10(10):591–596, 1979.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[11] S. Doyle, M. Feldman, J. Tomaszewski, and A. Madabhushi. A boosted bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies. *IEEE transactions on biomedical engineering*, 59(5):1205–1218, 2012.

[12] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017.

[13] A. B. Farris, C. D. Adams, N. Brousaides, P. A. Della Pelle, A. B. Collins, E. Moradi, R. N. Smith, P. C. Grimm, and R. B. Colvin. Morphometric and visual evaluation of fibrosis in renal biopsies. *Journal of the American Society of Nephrology*, 22(1):176–186, 2011.

[14] A. B. Farris and C. E. Alpers. What is the best way to measure renal fibrosis?: A pathologist's perspective. *Kidney international supplements*, 4(1):9–15, 2014.

[15] Y. Liu. Renal fibrosis: new insights into the pathogenesis and therapeutics. *Kidney international*, 69(2):213–217, 2006.

[16] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[17] A. Mahmood, M. Bennamoun, S. An, and F. Sohel. Resfeats: Residual network based features for image classification. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 1597–1601. IEEE, 2017.

[18] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer, 2011.

[19] P. Mathur, M. Ayyar, S. Chopra, S. Shahid, L. Mehnaz, and R. Shah. Identification of emergency blood donation request on twitter. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop and Shared Task*, pages 27–31, 2018.

[20] P. Mathur, R. Sawhney, M. Ayyar, and R. Shah. Did you offend me? classification of offensive tweets in hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, 2018.

[21] P. Mathur, R. Shah, R. Sawhney, and D. Mahata. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, 2018.

[22] H. Nejati, H. A. Ghazijahani, M. Abdollahzadeh, T. Malekzadeh, N.-M. Cheung, K. H. Lee, and L. L. Low. Fine-grained wound tissue analysis using deep neural network. *arXiv preprint arXiv:1802.10426*, 2018.

[23] Y. Peng, X. He, and J. Zhao. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing*, 27(3):1487–1500, 2018.

[24] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 512–519. IEEE, 2014.

[25] A. M. Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38(1):35–44, 2004.

[26] R. Sawhney, P. Manchanda, P. Mathur, R. Shah, and R. Singh. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 167–175, 2018.

[27] R. Sawhney, P. Manchanda, R. Singh, and S. Aggarwal. A computational approach to feature extraction for identification of suicidal ideation in tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 91–98, 2018.

[28] R. Sawhney, P. Mathur, and R. Shankar. A firefly algorithm based wrapper-penalty feature selection method for cancer diagnosis. In *International Conference on Computational Science and Its Applications*, pages 438–449. Springer, 2018.

[29] R. Sawhney, R. Shankar, and R. Jain. A comparative study of transfer functions in binary evolutionary algorithms for single objective optimization. In *International Symposium on Distributed Computing and Artificial Intelligence*, pages 27–35. Springer, 2018.

[30] L. Scarfe, A. Rak-Raszewska, S. Geraci, D. Darssan, J. Sharkey, J. Huang, N. C. Burton, D. Mason, P. Ranjzad, S. Kenny, et al. Measures of kidney function by minimally invasive techniques correlate with histological glomerular damage in scid mice with adriamycin-induced nephropathy. *Scientific reports*, 5:13601, 2015.

[31] R. Shah and R. Zimmermann. *Multimodal analysis of user-generated multimedia content*. Springer, 2017.

[32] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

Pre-print of paper accepted at WACV 2019

[33] K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206, 2016.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[35] P. D. Walker, T. Cavallo, and S. M. Bonsib. Practice guidelines for the renal biopsy. *Modern Pathology*, 17(12):1555, 2004.

[36] H. Wang, A. C. Roa, A. N. Basavanhally, H. L. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi. Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features. *Journal of Medical Imaging*, 1(3):034003, 2014.

[37] T. Wiatowski and H. Bölcskei. A mathematical theory of deep convolutional neural networks for feature extraction. *IEEE Transactions on Information Theory*, 64(3):1845–1866, 2018.

[38] S. C. Wong, A. Gatt, V. Stamatescu, and M. D. McDonnell. Understanding data augmentation for classification: when to warp? In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–6. IEEE, 2016.

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[40] M.-L. Zhang, J. M. Peña, and V. Robles. Feature selection for multi-label naive bayes classification. *Information Sciences*, 179(19):3218–3229, 2009.

Pre-print of paper accepted at WACV 2019