A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

http://wrap.warwick.ac.uk/169409

Copyright and reuse:
This thesis is made available online and is protected by original copyright. Please scroll down to view the document itself. Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk
Deep Learning Based Frameworks for Patient Selection

by

Ruqayya Awan

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Computer Science

June 2022
# Contents

List of Tables \hspace{1cm} iv  
List of Figures \hspace{1cm} vii  
Acknowledgments \hspace{1cm} xv  
Sponsorship and Grants \hspace{1cm} xvi  
Declarations \hspace{1cm} xvii  
Publications \hspace{1cm} xviii  
Abstract \hspace{1cm} xix  
Abbreviations \hspace{1cm} xx  

## Chapter 1 Introduction 1

1.1 Cancer \hspace{1cm} 1  
1.1.1 Cancer Screening \hspace{1cm} 1  
1.1.2 Tumour Heterogeneity \hspace{1cm} 2  
1.1.3 Types of Cancer \hspace{1cm} 3  
1.2 Cancer Diagnosis \hspace{1cm} 5  
1.2.1 Colorectal Cancer \hspace{1cm} 5  
1.2.2 Bladder Cancer \hspace{1cm} 6  
1.3 Digital and Computational Pathology \hspace{1cm} 8  
1.4 Thesis Aims \hspace{1cm} 9  
1.4.1 Patient Selection with High Risk for Bladder Malignancy \hspace{1cm} 10  
1.4.2 Whole Slide Image Registration \hspace{1cm} 11  
1.4.3 Patient Selection for Immune therapy \hspace{1cm} 13  
1.5 Thesis Organisation \hspace{1cm} 15  
1.6 Summary \hspace{1cm} 15
**Chapter 2  Identifying Patients with High Risk for Bladder Cancer**  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Introduction</td>
<td>16</td>
</tr>
<tr>
<td>2.2 Literature Review</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Materials and Methods</td>
<td>18</td>
</tr>
<tr>
<td>2.3.1 Specimen Collection, Digitisation and Labelled Data Preparation</td>
<td>19</td>
</tr>
<tr>
<td>2.3.2 ROI Extraction from Whole Slide Image</td>
<td>21</td>
</tr>
<tr>
<td>2.3.3 Cell Segmentation</td>
<td>22</td>
</tr>
<tr>
<td>2.3.4 Cell Detection and Classification</td>
<td>23</td>
</tr>
<tr>
<td>2.3.5 WSI-level Classification</td>
<td>25</td>
</tr>
<tr>
<td>2.4 Comparative Method</td>
<td>25</td>
</tr>
<tr>
<td>2.5 Experimental Results</td>
<td>26</td>
</tr>
<tr>
<td>2.5.1 Cell-level Classification</td>
<td>26</td>
</tr>
<tr>
<td>2.5.2 WSI-level Classification/Risk Assessment</td>
<td>26</td>
</tr>
<tr>
<td>2.5.3 Additional Experiments</td>
<td>27</td>
</tr>
<tr>
<td>2.6 Discussion</td>
<td>30</td>
</tr>
<tr>
<td>2.6.1 Annotation Variability</td>
<td>31</td>
</tr>
<tr>
<td>2.6.2 Performance of RetinaNet</td>
<td>31</td>
</tr>
<tr>
<td>2.6.3 Correlation between Cytology and Histology</td>
<td>32</td>
</tr>
<tr>
<td>2.6.4 Inter-observer variability w.r.t. WSI labelling</td>
<td>33</td>
</tr>
<tr>
<td>2.7 Summary</td>
<td>34</td>
</tr>
</tbody>
</table>

**Chapter 3  Handcrafted Feature Matching based Registration**  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Introduction</td>
<td>36</td>
</tr>
<tr>
<td>3.2 Literature Review</td>
<td>38</td>
</tr>
<tr>
<td>3.3 Hand-Crafted Feature-Based Registration</td>
<td>40</td>
</tr>
<tr>
<td>3.3.1 Maximally Stable Extremal Regions</td>
<td>41</td>
</tr>
<tr>
<td>3.3.2 Multi-scale MSER</td>
<td>43</td>
</tr>
<tr>
<td>3.3.3 Pre-processing steps</td>
<td>44</td>
</tr>
<tr>
<td>3.4 Datasets &amp; Performance Measures</td>
<td>47</td>
</tr>
<tr>
<td>3.4.1 Datasets</td>
<td>47</td>
</tr>
<tr>
<td>3.4.2 Performance Measures</td>
<td>48</td>
</tr>
<tr>
<td>3.5 Experimental Results</td>
<td>50</td>
</tr>
<tr>
<td>3.5.1 Experimental Setup</td>
<td>50</td>
</tr>
<tr>
<td>3.5.2 Results</td>
<td>50</td>
</tr>
<tr>
<td>3.6 Summary/Discussion</td>
<td>53</td>
</tr>
</tbody>
</table>

**Chapter 4  Deep Feature Matching based Registration**  

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 Introduction</td>
<td>54</td>
</tr>
<tr>
<td>4.2 Literature Review</td>
<td>55</td>
</tr>
</tbody>
</table>
### Chapter 4 Assessment of the DNA Mismatch Repair Status from Histology Images

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>80</td>
</tr>
<tr>
<td>Literature Review</td>
<td>83</td>
</tr>
<tr>
<td>MMR Protein Interpretation</td>
<td>84</td>
</tr>
<tr>
<td>The Proposed Method</td>
<td>86</td>
</tr>
<tr>
<td>5.4.1 Tumour Epithelial Segmentation</td>
<td>87</td>
</tr>
<tr>
<td>5.4.2 MSI/MSS Classification</td>
<td>90</td>
</tr>
<tr>
<td>Dataset and Performance Measures</td>
<td>92</td>
</tr>
<tr>
<td>Experimental Results</td>
<td>95</td>
</tr>
<tr>
<td>5.6.1 MMR Prediction</td>
<td>95</td>
</tr>
<tr>
<td>5.6.2 MSI Prediction</td>
<td>97</td>
</tr>
<tr>
<td>5.6.3 Comparative Analysis</td>
<td>99</td>
</tr>
<tr>
<td>5.6.4 TCGA Cohort</td>
<td>100</td>
</tr>
<tr>
<td>Summary/Discussion</td>
<td>101</td>
</tr>
</tbody>
</table>

### Chapter 5 Conclusions and Future Directions

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Identifying Patients with High Risk for Bladder Cancer</td>
<td>107</td>
</tr>
<tr>
<td>6.2 Registration of multi-stained WSIs</td>
<td>109</td>
</tr>
<tr>
<td>6.3 Computerised Assessment of Mismatch Repair Status for MSI Prediction</td>
<td>110</td>
</tr>
<tr>
<td>6.4 Concluding Remarks</td>
<td>111</td>
</tr>
</tbody>
</table>
## List of Tables

2.1 Total number of annotations marked by the pathologists and the sample split we used to train our initial and the final network. Further annotations for challenging cells were added in an iterative way. ................................................................. 20

2.2 Comparison of different CNN based classifiers on our validation set. Note that for the methods presented in the first five rows, classification was performed using the threshold-based segmentation maps while the last row presents the evaluation results of a one-stage detector. ................................................................. 26

2.3 Table presenting intra-observer variability of cell-level annotations provided in two attempts (Att# 1 and 2) by a pathologist. The Cohen’s Kappa is 0.15 showing a slight agreement for these classes; mainly the disagreement is in between atypical and malignant classes and in between atypical and others classes. . 31

2.4 Confusion matrix for manual cytopathology based risk vs manual histopathology based diagnosis (Sensitivity = 0.47, Specificity = 0.56). ................................................................. 33

2.5 Table presenting inter-observer variability in labelling the WSIs. Considering the two class classification (low and high risk), the agreement was found on 5 cases only out of 20 cases between P-1 (Reference) and P-2 (Pathologist-A). ......................... 35

3.1 A summary of methods used in the ANHIR challenge by the seven participating teams; NGF (normalised gradient field), NCC (normalised correlation coefficient), MI (mutual information), L-BFGS (limited memory Broyden-Fletcher-Globfarg-Shanno)................................................................. 40
3.2 Demonstration of improvement in registration error on applying feature matching at different resolutions in a multi-scale approach. The feature matching is applied for a WSI downsampled to a resolution level 7 for global transform. While for tissue and block-wise transforms, it is performed for the tissue region only at a resolution level of 6 and 5, respectively. Due to its large size, the tissue region is divided into four blocks at resolution 5.

3.3 Comparative results of registration using a single-scale, multi-scale MSER (MMSER) and case-wide approach. This comparative analysis is presented for two different settings: IHC vs IHC and H&E vs IHC. Registering IHC stained slide to H&E slide is challenging due to the differences in tissue architectures.

4.1 Comparative results of our proposed rigid alignment approach (DFBR) using two different tissue segmentation masks. TS1 refers to an overall tissue region mask including fatty tissue and TS2 refers to an active tissue region mask excluding fatty tissue. The exclusion of fatty areas significantly reduced the registration error. These results are generated using the COMET dataset.

4.2 Demonstration of the efficacy of our DFBR approach using different versions of input image pairs. Greyscale images are shown to outperform other pre-processed input images. These results are generated using the COMET dataset.

4.3 Details of the ANHIR dataset.

4.4 Parameters of non-rigid registration method: $\alpha$ is a regularisation parameter to control the smoothness in deformation while $\varepsilon$ controls the sensitivity of NGF to the noise, and * indicates the parameter that differs in the COMET dataset. We performed registration at a total of 6 levels for the COMET dataset.

4.5 Quantitative results generated using the COMET dataset. These results demonstrate that the average error is reduced with each alignment step in our pipeline.

4.6 Quantitative results of the winning team and our pipeline on the ANHIR dataset. First row in the header represents the aggregation method for an image pair and second row represents the aggregation method for all pairs in a set.
4.7 Comparative results of registration using a single-scale, multi-scale MSER (MMSER) and case-wide approach on the COMET dataset. This comparative analysis is presented for two different settings: IHC vs IHC and H&E vs IHC. Registering IHC stained slide to H&E slide is challenging due to the differences in tissue architectures.

5.1 MMR ground truth labels for the COMET dataset.

5.2 Average DICE values computed for each MMR marker across 5-fold cross-validation.

5.3 Average AUROC values for MSI prediction using 5-fold cross-validation obtained with different kernel functions in SVM.
List of Figures

1.1 An illustration of steps for preparation of histology slides.

1.2 Resected colorectal whole slide images taken from a single tissue block. The sections are stained with: a) CK8/18, b) E-cadherin, c) EpCAM, d) H&E, e) KI67, f) MLH1, g) MSH2, h) MSH6, i) P53, j) PMS2, k) PTEN and l) Vimentin.

1.3 An example of a urine cytology sample.

1.4 A high-level illustration of the proposed pipeline for risk stratification.

1.5 An example of two serial sections before and after registration. The image on left is a reference image stained with CK8/18 while the image on right is a H&E stained moving image.

1.6 A high-level illustration of our proposed pipeline for MSI prediction.

2.1 The illustration of our iterative process of collecting training data and performing qualitative testing at the same time using WSIs.

2.2 Some example images of synthetic cells generated using SMOTE.

2.3 The illustration of our proposed method for cell detection and classification from a WSI. (a) ROI detection (b) patches of size 5000 × 5000 pixels are extracted from ROI (c) unit which process every patch and output the coordinates and predicted label of each candidate cell (c1) cell segmentation followed by connected component analysis (c2) patch extraction (c3) classification of all the candidate cells into one of the seven classes using a trained classifier.

2.4 Result of cell segmentation. Note that the left image is a synthetic image generated using the cell patches extracted from the WSI. The right side image shows the segmentation map (in green) overlaid on the original image.
2.5 The Xception architecture. Each standard convolution and separable convolution is applied with a window size of $3 \times 3$ and is followed by batch normalisation. ........................................ 24

2.6 Results of cell classification using the Xception network. The left column shows the confusion matrix and the right column shows the ROC for binary classification. For binary cell classification, malignant and atypical cells are considered in a positive class while all the other classes are kept in a negative class. ........ 27

2.7 Supervised UMAP projection of the count of all seven types of cells predicted by the Xception network. Plus and dot markers are used for high-risk and low-risk cases respectively. ........... 28

2.8 Demonstrating the significance of the count of atypical and malignant cells for binary classification (low-risk vs high-risk) at WSI level. a) ROC curves using a count of malignant cells (M) and b) ROC curves using a total count of atypical and malignant cells (A+M). ROC is shown with 5-cross validation and the average area under curve using M and A+M is 0.81 and 0.83 respectively. c) A violin plot displaying the count of predicted atypical and malignant cells in WSIs belonging to low and high-risk classes. To generate these results, cells were classified using the Xception network. ......................... 29

2.9 Scatter plot demonstrating the relation between the histology-based (Benign and Malignant) and cytology-based (CA and ASM) diagnosis. Each marker represents a case. The urothelial cells were classified using the Xception network. ................. 33

2.10 ROC curves for Digital Cytology based risk vs Manual Histopathology based diagnosis. The sensitivity of all the predictors is computed for a fixed corresponding specificity of 0.56. To determine the digital cytology based risk, urothelial cells were classified using the Xception network. The confusion matrix for manual cytopathology based risk vs manual histopathology based diagnosis is provided in the Table 2.4. ......................... 34

2.11 A Venn diagram presenting inter-observer variability in labelling the 20 WSIs. These cases were misclassified by the method w.r.t. the reference but 15 out of those predictions agreed with another pathologist-A. The term reference represents the diagnostic information obtained from the UHCW while pathologist-A is an independent pathologist from a different hospital. To generate these results, cells were classified using the Xception network. 35
3.1 A block diagram showing the overall flow of the single-scale MSER based registration approach. ....... 42
3.2 The input images and architecture of the CNN model used for tissue segmentation. a) H stain image downscaled at magnification level 5, b) ground truth tissue mask and c) CNN model. The colour blue, green and red shown in b) represent background, slide artefacts and tissue region, respectively. ... 45
3.3 An example of input image preprocessing steps. The variance of the reference H channel is greater than that of the moving H channel. Therefore, during the normalisation process, the H channel of the moving image is modified such that its histogram is similar to that of the reference H channel. A tissue mask was used to exclude the non-tissue area while constructing the histogram. .................. 46
3.4 An example to illustrate the exact sequence of sections cut from a tissue block in the COMET dataset. .................. 48
3.5 Similarities and dissimilarities in the tissue architecture among spatially corresponding visual fields extracted from a) CK8/18, b) MLH1 and c) H&E stained images. a and b are taken from the consecutive slides (5µm apart), hence has more similarities as compared to a and c or b and c (more than 50µm apart). . 48
3.6 Manually annotated landmarks shown on corresponding tissue regions cropped from MLH1, MSH2, MSH6 and PMS2 stained slide images (left to right). ............ 50
3.7 Bar charts demonstrating registration efficacy and processing time for each case individually. Initial refers to the measures computed before alignment while global, tissue and block-wise are the series of transforms proposed for multi-scale MSER based registration (as discussed in section 3.3.3). The top and middle bar charts present registration performance for different steps in our multi-scale approach. Similarly, the comparison of processing time is shown in the bottom row, where single-scale corresponds to [2]. .......................... 52
4.1 Overall pipeline for our proposed cross-slide image registration. 58
4.2 An example image with the two types of tissue masks. In TS_1, all the tissue including fatty region is included whereas in TS_2, fatty region is excluded due to its negative impact on our DFBR registration method. .......................... 60
4.3 A qualitative comparison of the registration accuracy of transformations estimated using segmentation TS_1 and TS_2. The first and second columns in the first two rows show reference and moving images, respectively while the third and last columns show transformed moving images using segmentation TS_1 and TS_2, respectively. i) shows overlaid images of e) and g) while j) shows overlaid images of e) and h). Landmarks are overlaid for the purpose of visualising the offset between the two images.

4.4 A partial architecture of the VGG16 model, used for feature extraction in our DFBR module.

4.5 Demonstration of feature summation for a sub-image.

4.6 Demonstration of how quality of matched features is determined. An example distance matrix computed between the CNN features of two images is shown. Features with the smallest distance $\min_1$ are considered as matched features. $\min_2$ is the second smallest distance after $\min_1$. The quality of matched features is determined by the difference between $\min_1$ and $\min_2$.

4.7 Example images of the local phase-based refinement method integrated into our visualisation tool (see section 4.4) to fix translation offset. In bottom row g) we show overlaid false colour images before (left) and after refinement (right). Reference and registered moving images are shown in green and purple colours, respectively.

4.8 Cropped tissue region of a registered image overlaid on the corresponding region of a reference image with their ground truth landmarks. By looking at the landmarks, it is quite evident that some regions are not perfectly registered. This is due to the non-linear deformations in the tissue.

4.9 Visualisation Tool. Reference and registered moving images are displayed on the left and right side panels, respectively. Zooming and panning on these panels are interlinked. It allows the user to visualise corresponding regions of two registered WSIs at any resolution level.

4.10 Example images from the ANHIR dataset. Each image is stained with a different stain.
4.11 Quantitative results of our proposed framework on the COMET dataset. These results are presented for all fifteen pairs of images used for evaluation. A bar plot is presented for each reference image (H&E, CK8/18, MLH1, MSH2) separately, except MSH6. The registration is performed to align PMS2 w.r.t. MSH6 with 0.0025, 0.0026 and 0.0073 as \( \text{Avg(Med}(r\text{TRE})) \), \( \text{Med(Med}(r\text{TRE})) \) and \( \text{Avg(Max}(r\text{TRE})) \), respectively.

4.12 Demonstration of the median rTRE before and after applying each registration module using a box plot. The median rTRE is computed for the COMET dataset.

4.13 Demonstration of the efficacy of our proposed rigid registration in a step-wise manner using overlaid images of reference and registered moving images from the COMET dataset. The top row shows reference and pre-aligned moving images. The middle row shows the overlaid images of the corresponding tissue region: left, middle and right overlaid images are shown as the output of the pre-alignment, the tissue based alignment and the block-wise alignment, respectively. The bottom row shows the zoomed in overlaid patches. Reference and registered moving images are shown in green and purple colours, respectively.

4.14 Demonstration of the median rTRE before and after each registration module using a box plot. The median rTRE is computed for the ANHIR training set only.

5.1 Visual fields taken from three cases with different MMR staining patterns. The tumour region is highlighted with a red contour for the images in the first column while the bottom image contains all tumour. The border colour represents the MMR status where green and red are used for patches extracted from WSIs with intact and loss status respectively. Top row: intact status for all MMR markers; Middle row: intact status for MLH1 and PMS2 while loss status for MSH2 and MSH6; Bottom row: loss of MLH1 and PMS2 while the intact status for MSH6 and PMS2. Notice the nuclear staining of internal positive controls (lymphocytes and stromal cells).

5.2 A block diagram showing the overall flow for the a) training and b) testing of our 2-step MSI classification pipeline. In this study, we experimented with CK8/18 and H&E as target WSIs.
5.3 An example of tumour mask overlaid (in green) on the down-sampled H&E stained image. On the right side, zoomed version of the regions highlighted in red and blue boxes are shown. The top patch belongs to the normal tissue area with some tumour region at the bottom left while the bottom patch shows the tumour area.

5.4 Tumour mask for H&E and CK8/18. Tumour masks overlaid on H&E and CK8/18 are shown in the left and right panels, respectively. A tumour mask was originally generated for H&E using a trained CNN model. The tumour mask for CK8/18 was generated by transforming the H&E mask using a transformation matrix computed using the DFBR method proposed in Chapter 4.

5.5 Visual results of nuclei segmentation using HoVerNet. a) CK8/18 image, b) H stain channel extracted from a), and c) nuclei segmentation mask overlaid on b). Blue colour is used for outlining the boundary of epithelial cells. Red blocks are used to highlight the area where epithelial nuclei are missed by the model.

5.6 Epithelial mask generated by a trained CNN model. The top and bottom rows show overlaid images of CK8/18 and H&E images respectively. Images shown in the third column contain normal epithelium which was excluded using the respective tumour mask.

5.7 The proposed CNN architecture for MMR prediction.

5.8 Co-localised MMR Prediction using CK8/18 as a target image. a) Downsampled WSIs of MMR IHC sections registered to CK8/18 stained WSI. b) An example input image used for training a CNN. Its ground truth label is defined by thresholding H-score values computed for the corresponding patches from MMR slides. The corresponding MMR patches are shown in c). The colour of the borderline in c) shows the expression status for the given patch. Green represents that the patch is positively stained with DAB while red shows the absence of DAB stain. Patches showed in b) and c) are highlighted in red circles in a).

5.9 Overview of the proposed MSI prediction step. The MMR probability heatmaps and corresponding normalised histogram of probabilities are shown for both MSS and MSI cases.
5.10 MMR prediction against CK8/18. a-d) MMR stained slides: MLH1, MSH2, MSH6 and PMS2, e-f) respective probability heatmaps (generated by our multi-headed network) overlaid on registered MMR stained WSIs and the final predicted WSI label is MSI, i-l) zoomed in patches highlighted on a-h with small rectangles. The ground truth MMR expression status for a-d is Loss, Intact, Intact and Loss.  

5.11 Evaluation of our MMR prediction step using the DICE coefficient. The overlap is computed between the network generated MMR binary maps and ground truth H-Score maps after thresholding.  

5.12 Scatter plot showing the WSI-level MMR prediction results using our baseline experiment in comparison to that of the pathologist’s assignment. The corresponding confusion matrix is shown below each scatter plot. Any MMR slide with Intact status is given a value of 1 while a slide with Loss status is given 0. This plot was generated by taking the sum of these MMR values. The size of the circular markers is relative to the number of cases with matching pathologist and algorithm generated labels.  

5.13 MSI Prediction on CK8/18 and H&E images using localised and unlocalised settings. AUROC plots of 5-fold cross-validation along with average AUROC is shown. For each curve, x- and y-axis represent false positive and true positive rates.  

5.14 Example images from the TCGA cohort labelled as MSI. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has performed well for these images.  

5.15 Example images from the TCGA cohort labelled as MSI. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has not performed well for these images.  

5.16 Example images from the TCGA cohort labelled as MSS. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has performed well for these images.
5.17 Example images from the TCGA cohort labelled as MSS. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has not performed well for these images.

5.18 Comparison between TCGA and COMET cohorts in terms of the distribution of histological type (mucinous and non-mucinous) by MSI/MSS status.
Acknowledgments

Firstly, I'd like to express my special thanks to my supervisor, Prof Nasir Rajpoot for his guidance and feedback throughout my research work. I am extremely grateful for his invaluable supervision, support and patience during my PhD.

I would like to thank all the current and previous members of the TIAlab for all their support and suggestions throughout my PhD. Special thanks to Muhammad Shaban and Korsuk Sirinukunwattana, from whom I have learned the most. I am also grateful to Dr Fayyaz Minhas and Dr Shan E Ahmed Raza for their insightful comments and suggestions on my projects. I would also like to thank my friends and lab mates, especially Muhammad Shaban and Najah Alsubaie for a wonderful time spent together in the lab, at conferences and in social settings.

My gratitude extends to the clinical collaborators, Dr Ksenija Benes, Dr Clare Verrill, Dr Ayesha Azam, Dr Yee Wah Tsang, Dr Mohammed Nimir and Dr Heeba Maqbool, for all their work on data provision, creating labelled data and sharing their clinical knowledge. Special thanks to Dr Ksenija Benes for her hard work on marking annotations for the urine cytology project. Weekly face-to-face sessions with her helped me in understanding the cytology data and it was time well spent.

Finally, I must express my profound gratitude to my mother and husband. Completing this work would have been more difficult without their support and encouragement all through my studies. Last but not least, I am forever grateful and indebted to my father for all his support and confidence in me which encouraged me to pursue my research work overseas. I dedicate this work to the blessed memory of my beloved father, Muhammad Latif Awan.
Sponsorship and Grants

I would like to acknowledge the University of Warwick for giving me the opportunity and providing me with financial support in the form of the Chancellor’s International Scholarship.
Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. I declare that, except where acknowledged, the material presented in this thesis is my own work, and has not been previously submitted for obtaining an academic degree.

Ruqayya Awan
June 2022
Publications


Abstract

Recently, deep learning (DL) has become a spearhead for solving many problems in the computer vision domain, with computational pathology (CP) has no exception. In the CP domain, it is widely used for histological assessment of tissue for diagnosis and prognosis of cancer patients. The research community has developed an abundance of DL based CP tools, reporting state-of-the-art results, for many diverse applications. In near future, we can envisage better tools on the way forward to the clinical workflow to assist pathologists in making diagnostic and anti-cancer therapeutic decisions. In this thesis, we develop DL based frameworks defining eligibility criteria for selecting patients of two different types of cancers: bladder and colorectal cancers.

We develop our first framework with the main goal to investigate an automated alternative to risk stratification of urine cytology slides. We utilised digital cell profiles for the identification of patients with low-risk and high-risk of developing bladder cancer. Our experiments demonstrate that the digital risk could be a better predictor of the final histopathology based diagnosis. We then develop our second framework for the assessment of mismatch repair (MMR) status to identify patients with microsatellite instability (MSI), known to respond well to immunotherapy. We perform multi-stain tissue analysis using slides stained for MMR protein, in addition to H&E and cytokeratin stained slides. To the best of our knowledge, it is the first time that MMR status is utilised for MSI prediction. Registration is an important pre-requisite task before this multi-stain slide analysis. To this end, we present two approaches utilising two different features, hand-crafted and data-driven features. We adopted a multi-scale and multi-stage strategy, important for improving the quality of registration. These methods are able to align the images with low registration error as compared to other hand-crafted based approaches.
Abbreviations

**AUC** : Area Under the Curve  
**ANHIR** : Automatic Non-rigid Histological Image Registration  
**AMrTRE** : Average of Median Normalised Target Registration Error  
**AMaxrTRE** : Average of Maximum Normalised Target Registration Error  
**ASM** : Atypia Suspicious for Malignancy  
**BC** : Bladder Cancer  
**COM** : Centre of Mass  
**CNNs** : Convolutional Neural Networks  
**CRA** : Colorectal Adenocarcinoma  
**CRC** : Colorectal Cancer  
**CK8** : Cytokeratin 8  
**CSS** : Curvature Scale sSpace  
**CA** : Cytological Atypia  
**DFBR** : Deep Feature based Registration  
**dMMR** : Deficient Mismatch Repair  
**DP** : Digital Pathology  
**DAB** : Diaminobenzidine  
**FOV** : Field of View  
**FDA** : Food and Drug Administration  
**GU** : Genitourinary  
**GI** : Genitourinary  
**GT** : Ground Truth  
**H&E** : Haemotoxylin and Eosin  
**H** : Haemotoxylin  
**HNPPC** : Hereditary non-Polyposis Colorectal Cancer  
**HGUC** : High-Grade Urothelial Cancer
IHC : Immunochemistry
IDaRS : Iterative Draw-and-Rank Sampling
LBC : Liquid-based Cytology
L-BFGS : Limited Memory Broyden-Fletcher-Globfarg-Shanno
MMR : Mismatch Repair
MSI : Microsatellite Instability
MSI-H : Microsatellite Instability High
MSS : Microsatellite Stable
MSER : Maximally Stable Extremal Regions
ML : Machine Learning
MOPs : Multi-scale Oriented Patches
MMrTRE : Median of Median Normalised Target Registration Error
MLH1 : MutL Homologue 1
MSH2 : MutS Homologue 2
MSH6 : MutS Homologue 6
MI : Mutual Information
NGF : Normalised Gradient Fields
NCC : Normalised Correlation Coefficient
PMS2 : Postmeiotic Segregation Increased 2
pMMR : Proficient Mismatch Repair
PCR : Polymerase Chain Reaction
ROC : Receiver Operating Characteristic
ROI : Regions of Interest
SVM : Support Vector Machine
SMOTE : Synthetic Minority Over-sampling Technique
SIFT : Scale Invariant Feature Transform
TPS : The Paris System
TCC : Transitional Cell Carcinoma
TRE : Target Registration Error
TILs : Tumour-infiltrating lymphocytes
TCGA : The Cancer Genome Atlas
UHCW : University Hospitals Coventry and Warwickshire
WSIs : Whole Slide Images
Chapter 1

Introduction

1.1 Cancer

Cancer is a genetic disease characterised by excessive cell growth that happens due to mutations in the DNA. DNA is made up of genes containing all the instructions for a cell, giving signals to the cell for division or to stop cell division and repair any faults in other genes or cause cell death if the damage is non-repairable. In our body, normal cells are persistently undergoing division and death via a programmed process. Cancer happens when something goes wrong during the cell division, damage to DNA resulting in missing genes, multiple copies of genes, changes in the sequence of genes etc. These mutations deprive the cell of understanding the instructions and can accelerate abnormal cell multiplication. The uncontrollable growth of cells can form a mass called a tumour which is observed in two forms: benign or malignant. Benign tumours are non-cancerous and thus do not evade their primary place or spread to other parts of the body. On the other hand, malignant tumours are cancerous and can invade other parts of the body via a process known as metastasis. In general, cancer is a vast topic, hence in this chapter, we will highlight only those topics which are more relevant to this thesis.

1.1.1 Cancer Screening

Cancer is the most common cause of death worldwide, with 10 million deaths reported in 2020 [3]. According to a study [4], reporting the estimated rate of incidence and death due to cancer for the period 2014-2035, there would be an inflation of both death and incidence rates. In a recent study [5], there is an estimated 20% lifetime risk that a person will develop cancer and a 10% risk of dying due to cancer. The health systems globally, particularly in developing countries, are not prepared to deal with the increasing cancer burden. The lack of timely diagnosis and treatment leads to increased suffering and high mortality risk for cancer patients. This scenario can be avoided and
significantly controlled by improving the early diagnosis of cancer since cancer detected at its early stage is more likely to respond well to less expensive anti-cancer therapy resulting in a better survival rate for cancer patients. It can also significantly reduce the delay in patient care resulting in improvement in the overall lives of the patients.

Screening strategies are also in practice to combat the rapidly increasing incidence rate. Cancer screening refers to assessing a healthy individual without any symptoms of cancer and is different from early diagnosis which seeks to identify the individual with general signs and symptoms of cancer. Certain cancers can be avoided with screening such as during bowel screening (colonoscopy), polyps can be removed before it develops into cancer in the first place. While there are some general screening programs such as cervical screening programs which invite women in an age group of 25 to 64 years, there are certain screening tests that are only recommended for a particular group of people where genetic testing is used to look for Lynch syndrome in people with a family history of colorectal cancer. Other examples of screening methods are the PAP smear test for cervical cancer, screening mammograms for breast cancer and urine cytology for bladder cancer.

An abnormal result doesn’t confirm the disease. In general, further tests are conducted for confirmation. Most commonly, a tissue sample taken from the suspected body region is examined under a microscope. The microscopic assessment of the extent of cellular changes assists the physicians in determining the best treatment plan for the patient.

1.1.2 Tumour Heterogeneity

Gene mutations can occur in cells of any organ and hence cancer can develop anywhere in the body. The starting point for all types of cancer is the high proliferation of cells; however, they all behave differently due to varying molecular makeup and have their unique features, risk factors, survival rates and hence demanding different treatment and management plans. In addition, varying histological appearances within an organ is also not uncommon and their assessment helps in elucidating different sub-types of cancer and hence serves as a biomarker in diagnosis and prognosis. Apart from spatial heterogeneity across different tissue organs, temporal heterogeneity is also observed during the course of cancer due to changes in molecular signatures of the tumour. Patients with more heterogeneous cancer are likely to respond well to some therapies while being less sensitive to other anti-cancer treatments [6]. Tumour heterogeneity can also occur due to the treatment, demanding a dynamic approach for adjusting treatment to the dynamic disease. Therefore, an accurate judgement of tumour heterogeneity is crucial for making effective treatment plans.
Most of the malignant transformations are due to somatic mutations which are non-hereditary and are caused by the lifestyle choices like smoking, diet, lack of exercise etc. Although rare, they can be inherited from parents (germline mutation) and those people with mutated genes are at high risk of developing cancer. The advancement in molecular testing and comprehensive genomic profiling in clinical oncology has given an insight into the molecular makeup of the tumour and has been a stepping stone towards identifying personalised curable anticancer treatments. One example of germline mutations is Hereditary non-polyposis colorectal cancer (HNPCC), also known as the Lynch syndrome results from a genetic defect affecting the DNA mismatch repair (MMR) system. There are several methods for the identification of these mutations. There are two most commonly used methods: genetic analysis with a polymerase chain reaction (PCR) test and immunohistochemistry (IHC). PCR testing is considered to be a gold standard for MSI testing however IHC presents a simple method for identifying MMR pathway protein and indirectly reflects the MSI occurrence [7]. Most commonly used anti-cancer treatments include surgery, chemotherapy and radiotherapy; however, cancer patients deficient in MMR proteins have been shown to respond well to immunotherapy [8]. Thus, a comprehensive understanding and identification of MMR deficiency are necessary for making treatment eligibility criteria for patient selection.

1.1.3 Types of Cancer

In this thesis, we investigate the potential of deep learning for identifying two different types of cancers. Broadly speaking, we develop approaches for selecting patients who may be likely to develop or are suffering from either genitourinary (GU) cancer and those diagnosed with lower gastrointestinal (GI) cancer. GU refers to the genital and urinary systems while lower GI is pertaining to the tract from the large bowel to the anus. In this section, we will discuss the cancer of specific organs falling under these two broad categories.

Genitourinary Cancer

GU cancers are male-dominant cancers. Among GU cancers, prostate cancer is the most common type of malignancy, followed by bladder and kidney cancers [9]. There are other less commonly diagnosed cancers under this category such as testicular, penile, adrenal and urethral cancers. In this thesis, we focus on algorithms for identifying patients with a high risk of bladder cancer from urine cytology often used for bladder cancer screening.

Bladder cancer (BC) is the 10th most common cancer globally, accounting for 3.2% of all cancers excluding non-melanoma skin cancer [10]. The rate of incidence of BC is increasing globally, particularly in developed countries. It is
ranked 6th and 17th for men and women, respectively, with respect to newly diagnosed cases in 2018, making it male-dominant cancer. The death rate due to this cancer is also observed to be four times higher in men than in women [11]. It is generally considered an elderly disease since the risk of developing BC increases with age, with 73 as the median age of diagnosis [12].

Bladder cancer (BC) can occur in urothelial cells which are present in the lining of the bladder as well as the tube connecting the bladder to the kidney and can spread to other parts of the body. Patients with BC may experience irritation, urine in the blood or changes in bladder habits. However, some patients may not experience any of these changes at their early stage. By the time any of these symptoms appear, cancer may have already metastasised, hence minimising the chances of survival.

Bladder cancer is one of the highly treatable and curable cancer if diagnosed early. Screening tests can play a vital role in early diagnosis and estimating the risk of malignancy, hence it can minimise the probability of dying from this cancer. Currently, there are no standard routine screening tests for bladder cancer. Urine cytology is recommended along with other tests to screen for cancer cells in urine samples. If abnormal cells are found in the urine sample, then further tests such as cystoscopy are recommended which involves looking inside the bladder. During cystoscopy, if any abnormal tissue is found, then a small tissue sample will be extracted for further examination at a microscopic level.

**Gastrointestinal Cancer**

GI cancers are causing more cancer-related deaths than any other cancer type worldwide, accounting for around 3.5 million deaths in 2020 [3]. The most commonly diagnosed cancers affecting the digestive system are colorectal cancer (CRC), gastric cancer, liver cancer, oesophagal cancer and pancreatic cancer [13]. Among them, CRC is the most commonly diagnosed cancer and among all the cancers, it is the third most common cancer in terms of newly reported cases and the second most common cause of cancer deaths for the year 2020 [4]. Unlike BC, CRC doesn’t exhibit gender discrepancy in its incidence rate and is diagnosed in both men and women equally worldwide. Similar to BC, CRC is also considered to be a disease in elderly people. However, recently there has been a surge of young patients with CRC [14, 15].

CRC develops when an abnormal growth, called a polyp, in the inner wall of the colon or rectum becomes malignant. The associated controllable risk factors include lifestyle choices such as diet, smoking, alcohol consumption, lack of exercise etc. However, there are 20% of patients who develop cancer due to hereditary conditions. There are four possible inherited conditions: Familial
adenomatous polyposis, MUTYH-associated polyposis, Gardner syndrome and hereditary nonpolyposis colorectal cancer (HNPCC). Among them, HNPCC also known as the Lynch syndrome, is the most common type of hereditary cancer [16]. Patients with the first three conditions are likely to have a large number of adenomas (benign polyp) and due to their abundance, there are more chances of their transitioning to malignancy. On the other hand, patients with the Lynch syndrome will have a normal number of adenomas as that of a normal person but these adenomas exhibit a higher tendency for transformation. Approximately 90% of CRC with Lynch syndrome are shown to have microsatellite instability (MSI) [16], a condition correlated with mutations in one of the genes responsible for the mismatch repair system.

CRC, when detected early, is also one of the cancers having a high cure rate after appropriate treatment. Endoscopy of the colon, called a colonoscopy, is used for both screening and early diagnostic purposes. It is recommended for people of age 50 or above to screen for CRC every 10 years or sooner. During this procedure for early diagnosis, if a doctor finds a tumour, a tissue sample is taken for microscopic examination. Based on the extent of malignancy and the results of other lab tests, the best treatment plan is recommended for the patient.

1.2 Cancer Diagnosis

In this thesis, our focus is on bladder and colorectal cancers, for which endoscopy and biopsy is the key diagnostic procedure. Endoscopy for the colon and bladder is called colonoscopy and cystoscopy respectively. It allows the physicians to look inside the organs. During this invasive procedure, a biopsy tissue sample is taken from the suspicious area of the organ for further assessment. The procedure of obtaining a sample depends on the location and type of the tumour. Below, we discuss sample types for each cancer that we used for patient selection (or diagnosis) in this work.

1.2.1 Colorectal Cancer

On encountering a suspicious polyp during colonoscopy, a tissue sample is taken for further microscopic assessment and on the basis of histological features, treatment is advised. For early stage colorectal cancer, a malignant polyp can be removed during colonoscopy as a part of the treatment. Whereas for more advanced colorectal cancer, treatment involves the surgical removal of the cancerous part of the colon or rectum, known as bowel resection. For our work on colorectal cancer, we have used resected sections acquired from the tumour during surgery. Figure 1.1 demonstrates different steps of a slide preparation
process from a processed tissue embedding or a tissue block (preserved tissue sample). Several consecutive sections may be cut from each tissue block and mounted onto microscopic slides for analysis. Each of these sections may then be stained with a specific biomarker for a detailed analysis. In this thesis, we present methods for the analysis of resection for our work on WSI registration (Chapters 3 and 4) and MSI prediction (Chapter 5). Examples of resected sections stained with different biomarkers are shown in Figure 1.2.

1.2.2 Bladder Cancer

The work in this thesis on bladder cancer is mainly based on cytological slides obtained for screening purposes. For screening, urine cytology is recommended which involves examining the cells from the urine samples under a microscope. Urine cytology is an important tool for identifying malignancy in the urinary tract and is mainly used for identifying high-grade urothelial cancer. For urine cytology, a sample can be obtained from normal urination or using a small tube. An example image of a urine cytology sample is shown in Figure 1.3. In clinical practice, cytology slides are observed under the microscope and based on cellular morphology, further tests are suggested.

The microscopic analysis of a histology slide is the gold standard for both low and high-grade cancer diagnoses. In comparison, urine cytology is used for high-grade cancer only. In this thesis, we also study the correlation between the cytopathology based diagnosis and histopathology based diagnosis. We obtained histopathology diagnoses for some cases along with their cytology and histology reports. The findings on histology reports were based on the biopsy slides obtained using a surgical procedure known as transurethral bladder tumour resection or TURBT. During this procedure, a tissue sample from the tumour region is removed based on the results of cystoscopy. A final diagnosis is made after a microscopic examination of the tissue sample.
Figure 1.2: Resected colorectal whole slide images taken from a single tissue block. The sections are stained with: a) CK8/18, b) E-cadherin, c) EpCAM, d) H&E, e) KI67, f) MLH1, g) MSH2, h) MSH6, i) P53, j) PMS2, k) PTEN and l) Vimentin.
1.3 Digital and Computational Pathology

With the advent of digital scanners, the pathologist’s practice for diagnosis is transforming from visual microscopic analysis to digitised tissue analysis. Their introduction has led to the possibility of using digital slides for analysis regularly, with some laboratories opting to go fully digital. This has also encouraged the machine learning community to contribute to designing computer algorithms that can provide an objective assessment of the tissue architecture. However, there is a significant lag in the uptake of these algorithms in a clinical setting. Recently, an machine learning (ML) based software known as Paige Prostate is approved by the U.S. Food and Drug Administration (FDA) to assist the pathologist in detecting prostate cancer [17]. Similar computational tools are likely to be deployed in future to assist the pathologists in making diagnostic decisions for other cancer types. These tools are expected to improve the accuracy and consistency of diagnosis which can ultimately provide an effective tool for management decisions. This is particularly relevant to CRC and BC due to the subjective nature of assessment. These tools are also likely to improve the turnaround time in a clinical practice which will result in a better overall outcome for the patients.

In general, digital pathology (DP) along with computational pathology has been successful in altering the pathology ecosystem. However, its usage for routine diagnostic work on cytopathology is rare, with occasional labs using...
it. Its widespread adoption has been hindered due to various challenges: lack of evidence around validation of the use of DP for cytology and technical and operational considerations where z-stacking is required. In practice, pathologists need to use the multi-plane focusing feature to look through the 3D clusters of cells frequently. While in current digitization practice, WSIs are commonly scanned at a single focal plane and assessing digital cytology slides is a challenge for pathologists. WSIs with z-stacking comprise a stack of images of a glass slide forming a composite multi-plane image. It has been demonstrated that the availability of more than one focal plane for cytopathology specimens helps with the diagnostic interpretation [18]. There are several scanners with z-stacking ability, scanning a glass slide at different focal planes along the vertical axis of the sample. However, the more images in a z-stacked WSI the greater the data storage requirement and the scanning time.

In recent years, researchers have proposed several computational methods that provide objective measures for a wide variety of tasks. Recently, due to the increasing computational power, deep learning based methods have made impressive strides as a result of their outperforming outcomes in several domains including the healthcare domain. Broadly speaking, it has been mainly used in three ways for cancer detection: 1) for getting the final output from the model, 2) to get intermediate results and 3) to learn representative features that could be used for several tasks. In our work, we have mainly used deep learning to generate intermediate results: we identify cases with a high risk of bladder malignancy based on the count of abnormal cells detected through deep learning and compute deep features for geometric transformation for MMR and MSI prediction.

1.4 Thesis Aims

The central aim of this thesis is to improve certain diagnostic decision-making that can assist with treatment planning for cancer patients. We employ different analytical strategies for bladder and colorectal cancers due to the differences in their respective samples for analysis. We perform cell-level analysis to identify patients with a high risk of BC whereas, for CRC, we perform multi-stain tissue-based analysis for identifying patients who can respond well to the immunotherapy. We undertake an additional step of registering whole slide images (WSIs) before multi-IHC section analysis. Below, we discuss the motivation, objective and innovation of our proposed methods for each cancer type.
1.4.1 Patient Selection with High Risk for Bladder Malignancy

Motivation

The prognosis of bladder cancer is dependent on the tumour size and how far it has spread which is collectively referred to as the stage of cancer. Patients with lower stage are observed to have a better overall outcome and any delays in the diagnosis can adversely affect the patient survival. Bladder cancer can often be diagnosed early in symptomatic patients. However, it can also cause no symptoms in the early stages and when any urinary symptoms appear, there is a high probability that cancer has already metastasised. Therefore, there is an obvious need for early detection of bladder cancer, for which arranging screening programs is one of the possibilities. Currently, urine cytology is used for symptomatic patients and can serve as a screening test for asymptomatic people. Visual assessment of cytology slides is subjective with large inter- and intra-observer variability [19]. Besides, screening programs increase the workload for clinicians. Computer-assisted screening program can help streamline the workload of clinicians and can provide a more objective diagnosis.

Objectives

The main goal of this strand of the thesis is to investigate an automated framework for identifying patients with a high risk of bladder malignancy using urine cytology slides. A high-level illustration of our proposed pipeline for risk stratification is shown in Figure 1.4. The developed framework has two main aspects. Firstly, it detects and classifies each candidate cell or group of cells in a slide. Automatic detection of diagnostically important cells can assist the pathologists in locating the highly suspicious cells in a slide, avoiding the need to scan through the whole slide and would improve the pathologist’s turnaround time. The second approach involves utilising a cell count based approach for identifying high-risk cases. A higher count of diagnostically important cells can help with recommending gold standard diagnostic procedures for high needs patients which can ultimately avoid delays in diagnosis and treatment and improve the overall patient outcomes.

Our Contributions

Our contributions to this work are five-fold. First, we iteratively collect cell-level annotations to improve the generalisability of the model. The TPS criteria, which are considered a reliable diagnostic tool, were used by the expert pathologists for labelling. Second, we explore two different approaches for cell detection and classification and employed the best one for WSI labelling. Third,
we present a cell count based approach for identifying high-risk cases. Fourth, we investigate the inter-observer agreement at the WSI level and intra-observer variability at the cell level. Lastly, we investigate the cytopathology based risk category and our digital risk labelling in relation to the ‘gold standard’ histopathology based diagnosis.

1.4.2 Whole Slide Image Registration

Motivation

In diagnostic and research practice, cross-slide image analysis provides additional information by analysing the expression of different biomarkers as compared to a single slide analysis. Most frequently, it is used to analyse adjacent sections stained with different biomarker stains, where, different stains are used to label different nuclear and tissue structures. Slides stained with different biomarkers are analysed side by side which may reveal unknown relations between the different biomarkers. During the slide preparation, a tissue section may be placed at an arbitrary orientation as compared to other sections of the same tissue block (see Figure 1.2). The problem is compounded by the fact that tissue contents are likely to change from one section to the next and there may be unique artefacts on some of the slides. This makes registration of each section to a reference section of the same tissue block an important pre-requisite task before any cross-slide analysis. Currently, this registration is done manually by the pathologists which is time-consuming due to the large number of sections taken from a single tissue block. Accurate registration is needed for any automated cross-slide analysis framework.

Objectives

We perform the registration of multi-stain resected sections of colorectal tissue. The end objective of registration is MSI prediction, involving analysis of MMR staining in the same part of the tissue as that of CK8/18 or H&E
image. The localised assessment of the staining pattern requires the geometric transformation of a WSI to ensure that corresponding tissue structures are at the same coordinate positions. An example pair of images before and after registration is shown in Figure 1.5. We propose two different approaches for the automatic registration of slides as a replacement for the manual approach. In Chapters 3 and 4, the registration problem is dealt with two different types of features: hand-crafted and data-driven features. We employ these features to perform a rigid transformation, followed by a non-rigid transformation to deal with non-linear tissue deformations. We adopted multi-scale and multi-stage strategies for hand-crafted and data-driven approaches for registration, respectively. These strategies are important for improving the quality of registration.

![Figure 1.5: An example of two serial sections before and after registration. The image on left is a reference image stained with CK8/18 while the image on right is a H&E stained moving image.](image_url)
Our Contributions

To the best of our knowledge, it is the first time that deep features are used for the estimation of rigid transformation for histology images. Our proposed method is able to align the images with low registration error as compared to other hand-crafted based approaches. To further mitigate the non-linear deformations, a non-rigid transformation can be applied which can only be optimal when two images are linearly aligned. Otherwise, it can introduce artificial artefacts in the transformed image. To evaluate the efficacy of our deep feature-based registration (DFBR) method, we experimented with the framework that won the Automatic Non-rigid Histological Image Registration (ANHIR) challenge. This framework consists of pre-alignment, parametric and non-parametric steps. We replaced the first two steps with our DFBR to register challenge provided image pairs. The modified framework produced comparable results to that of challenge winning team. We also developed a visualisation tool to view registered reference and source WSIs at different magnification levels. With the help of this tool, one can apply a transformation on the fly without the need to generate transformed source WSI in a pyramidal form.

1.4.3 Patient Selection for Immune therapy

Motivation

An accurate diagnosis and profiling of tumours are critical to the best treatment choices for cancer patients. A cautious decision needs to be taken while considering the risks and benefits of each therapy. There are many types of anti-cancer treatments and in addition to the cancer type and its aggressiveness, molecular heterogeneity also plays a vital role in treatment selection. MSI or MMR deficiency is one of the well-studied aberrations in terms of molecular changes and is observed in 15% of all colorectal cancers [20, 21]. CRC patients with MMR deficiency do not get benefits from chemotherapeutics as compared to CRC patients without MSI; instead, they respond well to immunotherapy [22]. Assessment of these molecular markers can assist clinicians in making optimal treatment selections for CRC patients. These molecular changes can be identified using IHC or PCR tests. Currently, these tests are only conducted for a subset of CRC patients as it incurs increased workload and high cost. The introduction of automated methods that can predict MSI or MMR status from a target image (H&E image in previous studies) without the need for additional sections of tumour tissue or genetic analysis can substantially reduce the cost associated with it.
Objectives

Following on from registration aligning MMR images to the reference image, we develop a framework for MSI prediction. We predict MSI status based on our prediction of MMR protein expressions. We employ deep learning to perform patch-based MMR prediction directly from a target histology image. To this end, we first obtain the expression status of all four MMR biomarkers by presenting MMR IHC WSIs (MLH1, MSH2, MSH6 and PMS2) along with the corresponding H&E WSI to a pathologist. We design a multi-headed convolutional neural network (CNN) model where each head is responsible for predicting one of the MMR protein expressions. We generate labelled data using stain colocalisation to train the model by assigning each patch extracted from the tumour epithelial region of the target image the diaminobenzidine (DAB) intensity of the corresponding MMR patches to assign the patch a binary expression label. Statistical features computed from the MMR prediction maps are then used for the final MSI prediction. A brief overview of our MSI prediction pipeline is shown in Figure 1.6. The description of this method is provided in Chapter 5.

Our Contributions

To the best of our knowledge, it is the first time that MMR protein expression is utilised for MSI prediction. In previous studies on MSI, coarse-grain information (microsatellite stable or unstable) has been used for training a model. However, in our work, we take the benefit of the fine-grained labels which comprise different patterns of MMR expression. Our results demonstrate that MSI classification can be improved by training CNN with fine-grain labels. We
predicted MMR expression and MSI status using tissue slides stained with H&E and cytokeratin 8 (CK8) paired with cytokeratin 18 (CK18). The cytokeratin stains are referred to as CK8/18 in this thesis. We also investigated the generalisability of our model for MMR prediction on the TCGA dataset for colorectal cancer.

1.5 Thesis Organisation

This thesis comprises six chapters in total.

- In Chapter 1, we present an introduction to the cancer types that are relevant to our thesis, along with a summary of our contributions toward the diagnosis and management of these cancers.

- Chapter 2 presents deep learning based digital cell profiles for risk stratification of urine cytology images.

- Chapters 3 and 4 present registration frameworks for resected multi-IHC WSIs of CRC. In Chapter 3, we present our basic registration method using multi-scale based hand-crafted features while Chapter 4 presents our deep feature matching based registration method.

- Chapter 5 provides an application of the registration work, MSI prediction. In this chapter, we present a computerised assessment of MMR status for MSI Prediction.

- Chapter 6 presents a summary of our proposed methods, limitations and some potential future directions for each of the proposed methods.

1.6 Summary

In this chapter, we briefly discussed the topics that constitute the motivation for this thesis work. We began with a short explanation of the biology of cancer. We introduced some of the cancer-related concepts which are more relevant to this thesis. We presented a brief introduction to Digital and Computational Pathology. We then presented the overall aim of the thesis with a brief summary of each part of our thesis work including its motivation, objectives and contributions. We ended this chapter with a short summary of the overall organisation of the thesis.
Chapter 2

Identifying Patients with High Risk for Bladder Cancer

2.1 Introduction

Urine cytology is considered to be an important detection tool for identifying malignancies in the urinary tracts such as bladder cancer. It is widely used to identify high-grade urothelial cancer (HGUC) and is not encouraged to be used for low-grade carcinoma due to its low sensitivity to it. In clinical practice, pathologists observe cytology slides under the microscope and identify atypical and malignant cells. Based on the morphology of these cells, a diagnosis is made leading to decision making for treatment.

Unlike histology, the digital adoption for urine cytology has been impeded due to the lack of the scanner’s ability for z-stacking along with other limitations related to cytology. The tissue material for histology has a relatively uniform thickness whereas the cytology material is less evenly distributed with variable thickness of different cell clusters in a 3D configuration. For this reason, the pathologists would frequently need to focus on different planes to view all the cells. It has been demonstrated that the availability of more than one focal plane on digital cytology slides helps with the diagnostic interpretation [18]. Z-stacking enables the user to look at the sample at different focal planes which is a built-in property of the microscope. With the advancement in whole slide scanners, different vendors have started to provide imaging systems with an ability for z-stacking which has motivated the pathologists to scrutinise digital cytology in clinical practice. However, it comes with a cost of a much larger image file size and longer scanning time [23].

Similar to histology, diagnosis of cytology cases suffer from high inter- and intra-observer variability [19]. In addition to variability in the assessment of urine cytology, different terms for the same entities were being used at both individual pathologists and institutional levels. This led to the development
of The Paris System (TPS) to provide a consistent and reliable diagnostic tool. An international working group, comprising expert cytopathologists, urologists and surgical pathologists, provided criteria for reporting different diagnostic categories including recommendations for HGUC which is the main purpose of urine cytology. TPS was officially released in 2016 [24] and is now accepted worldwide. It has shown significant improvement in the assessment of urine cytology specimens with adequate precision for negative cases. However, studies [25–28] conducted on the inter-observer variability demonstrated poor inter-observer agreement for other categories. In [28], different distribution of categories was reported by five cytopathologists on reviewing 149 cases independently. The inter-observer variation makes the diagnosis of cytology samples challenging.

In a clinical setting, a cytology specimen is examined manually, under a microscope using a glass slide. Like histology samples, urine cytology slides can be visualised on a computer screen after digitisation which is used by occasional labs. The uptake of digital cytology can encourage the assisted assessment of specimens with computer-generated results. This would result in the emergence of quantitative algorithms for analysis, hence enabling the clinicians to obtain non-subjective and reproducible outcomes.

The main goal of this study is to investigate an automated alternative to risk stratification of urine cytology slides so that people with a high risk of malignancy can be identified. The status quo based on subjective visual analysis is prone to human error and has a large inter- and intra-observer variability. Therefore, there is a need to investigate its limitations w.r.t. the intrinsic difficulty of the problem in both diagnostic and technical terms.

### 2.2 Literature Review

In the literature, very few studies can be found on automatic analysis of cytology images in comparison to the work in histology image analysis. Recently, there has been some work on cell detection, classification and segmentation from cytology images. In [29], GoogleNet and AlexNet models have been used to distinguish between benign and malignant microscopic images of breast cytological samples obtained with the fine needle technique. The training and validation dataset was collected by extracting overlapping patches (comprising a number of cells) from ROIs selected by the pathologist. Zhang et al. [30] presented a simple CNN to classify cervical cells in a pap-smear cytology image without any prior cell segmentation. Their training set comprises patches of fixed size with a nucleus located in the centre of the patch. This means that the network was trained with patches containing partial cell content. In another study [31], a simple CNN is used to classify the cells in nasal cytology into one
of the seven classes. To train the network with patches containing whole cell content, they perform cell segmentation via the Otsu algorithm followed by morphological operation and watershed algorithm. To overcome the problem of unbalanced classes, they opted random majority under-sampling method. Wu et al. [32] employed AlexNet based network to identify different types of ovarian cancer from cytological images captured from different parts of the tissue sample. These images were then divided and resized into smaller patches for training.

One recent study [33], which integrates deep learning and morphometric approaches focuses on automating TPS for the analysis of urine cytology images. Deep learning is used to assign an atypia score to a given cell while a morphometric approach computes the nucleus to cytoplasmic ratio. They employed thresholding to segment cellular content, followed by connected component analysis for extracting cell patches. Based on their cell classification approach, they proposed a condensed grid format for an image reconstruction which is less cellular and smaller in size in comparison to the original image. The authors also illustrated the prediction of high-risk cases based on the cutoff for their employed cell morphological features. Sanghvi et al. [34], presented a deep learning-based pipeline for classifying urine cytology images into five TPS categories which can further be divided into low and high-risk classes. QuPath was used to detect cells in a WSI and a patch of fixed size was extracted from the centre. The authors employed both cell-level and slide-level features for WSI classification and validated it using a large cohort. To the best of our knowledge, [33, 34] are the only studies on risk stratification.

There has been some effort in separating the overlapping cells from both 1-plane and z-stacked cytology images and is not limited to [35], [36], [37], [38], [35] and [39]. In this study, we perform segmentation to extract both individual cells and clusters of cells. This is to ensure that the whole cell or a cluster is captured inside the bounding box. Therefore, separating the overlapping cells is not necessary for our approach.

2.3 Materials and Methods

Atypical and malignant cells are of interest to the pathologist among various types of cells and contaminants found in a urine cytology sample. To discriminate between low and high-risk WSIs, we first identify all the candidate atypical and malignant cells. For our experiments, we employed a deep learning-based approach for the identification of these cells.
2.3.1 Specimen Collection, Digitisation and Labelled Data Preparation

The cytology slides used in this study and the associated clinical data were obtained from the University Hospitals Coventry and Warwickshire (UHCW) NHS Trust in Coventry, UK. The dataset was provided after de-identification and informed consent was obtained from the patients. Each slide was labelled as Normal, Inflammatory, Cytological atypia (CA), Atypia suspicious for malignancy (ASM) or Transitional cell carcinoma (TCC). In this chapter, we use the term ‘reference’ for the diagnostic information obtained from the UHCW which does not necessarily mean that it was decided by a single pathologist. All the slides were prepared using a liquid based cytology method, ThinPrep and were scanned at 0.275 microns/pixel. The maximum resolution is 40×.

In total, we obtained 398 slides, comprising 243 Normal, 13 Inflammatory, 76 CA, 38 ASM and 28 TCC. These slides were scanned using an Omnyx VL120 scanner to form a multi-layered pyramid enabling the user to visualise the slide at different resolutions.

Creation of Labelled Dataset

We obtained cell-level annotations from an experienced pathologist and a recently trained pathologist. Both pathologists followed TPS criteria for labelling cells as normal, atypical or malignant urothelial cells. Other cell types present in urine (e.g. squamous, inflammatory etc.) were also annotated. Degenerated cells and cells that pathologists were uncertain about were annotated as ‘Others’. The variations in annotations affect the performance of a trained classifier. We did the inter-observer variability analysis between two pathologists to find out the highly concordant classes. A set of the same visual fields were presented to both pathologists for independent annotations. A high concordance score was observed in normal, squamous and inflammatory classes. Considering the variability in the rest of the classes, we expanded our labelled dataset by presenting different visual fields to them. We asked the trained pathologist to annotate samples of normal, squamous and inflammatory only while the experienced pathologist annotated all the classes.

For network training and validation, annotations were obtained on the WSIs via a web-based interface. The pathologists marked individual cells using a dot in the centre of the cell and for the cell clusters, pathologists draw polygons or rectangles around clusters. The annotations were obtained from pathologists at resolution level 40×. The details of the sample split to train our initial network are shown in Table 2.1. More annotations were added to this dataset while verifying the network predictions by the expert pathologist in an iterative manner, as shown in Figure 2.1. During the verification process,
747 normal, 2185 squamous, 2408 others, 2073 debris, 117 inflammatory, 279 atypical and 88 malignant cells were added to the training set while 163 normal, 544 squamous, 492 others and 511 debris cells were added to the validation set.

Figure 2.1: The illustration of our iterative process of collecting training data and performing qualitative testing at the same time using WSIs.

<table>
<thead>
<tr>
<th>Class</th>
<th>Total</th>
<th>Iter:1</th>
<th>Iter:N</th>
<th>Iter:1</th>
<th>Iter:N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>2,645</td>
<td>1,235</td>
<td>1,982</td>
<td>500</td>
<td>663</td>
</tr>
<tr>
<td>Squamous</td>
<td>4,133</td>
<td>904</td>
<td>3,089</td>
<td>500</td>
<td>1,044</td>
</tr>
<tr>
<td>Inflammatory</td>
<td>1,394</td>
<td>777</td>
<td>894</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Others</td>
<td>4,273</td>
<td>873</td>
<td>3,281</td>
<td>500</td>
<td>992</td>
</tr>
<tr>
<td>Atypia</td>
<td>5,337</td>
<td>4,558</td>
<td>4,837</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Malignant</td>
<td>3,305</td>
<td>2,717</td>
<td>2,805</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Debris</td>
<td>7,642</td>
<td>4,558</td>
<td>6,631</td>
<td>500</td>
<td>1,011</td>
</tr>
<tr>
<td>Total</td>
<td>28,729</td>
<td>15,622</td>
<td>23,519</td>
<td>3,500</td>
<td>5,210</td>
</tr>
</tbody>
</table>

Table 2.1: Total number of annotations marked by the pathologists and the sample split we used to train our initial and the final network. Further annotations for challenging cells were added in an iterative way.

Balancing of Labelled Dataset

The dataset obtained after initial annotations suffered unbalanced distribution. Due to unbalanced classes, a classifier does not tend to perform well for the minority classes as it does not get a sufficient look at them. To balance the distribution in the training set, we employed an oversampling technique known as Synthetic Minority Over-sampling Technique (SMOTE) [40]. In our initial dataset, it was atypia and debris class for which most of the annotations were obtained from the expert pathologist. For network training, we kept 500 samples per class in the validation set and the remaining samples in the training set. Except for the atypia class, all other classes were oversampled until the total number of patches per class including the original and oversampled patches was equal to the number of patches in the atypia class (4,558). Since
the dimensionality of patches is not constant due to the various sizes of the cell, we applied SMOTE on patches in batches, with each batch having patches of similar sizes.

The SMOTE technique generates new samples by manipulating the feature space by joining the line segments between each minority class sample and its $k$ nearest neighbours. This is done by computing the difference between the minority sample and its nearest neighbour, followed by multiplying the difference with any random number between 0 and 1. For our experiments, we selected 5 nearest neighbours for synthesis. Examples of images of synthetic cells generated by the SMOTE technique are shown in Figure 2.2.

Figure 2.2: Some example images of synthetic cells generated using SMOTE.

2.3.2 ROI Extraction from Whole Slide Image

In histology slides, a relatively small region of the slide contains tissue and to reduce the computation time, the tissue region is identified to avoid processing the background white region. To exclude the background region, thresholding can be used for histology images at a low resolution. Like histology images, in the urine WSIs, the cellular content is confined to a limited portion of the slide. However, unlike, histology images, thresholding at low resolution would omit cells in WSIs with a fewer number of cells.

In Figure 2.3 (a), an example of a urine cytology slide from our dataset is shown at a low-resolution level. The area inside the two fiducial marks (black curvy lines) contains cells while the remaining area is non-cellular. Hence, the region outside these two fiducials should be excluded from processing to reduce the computation time. To achieve this, we adopt the Otsu thresholding [41] which determines a threshold value by maximising the inter-class intensity variance. Specifically, we first convert the RGB images into a greyscale image and then an optimal threshold value is estimated using the Otsu algorithm. A number of other objects such as the text on the slide and other artefacts were identified using this threshold value. These were excluded based on the area-based threshold. The resulting mask for the region of interest is shown in Figure 2.3 (a) and it was carried out at a resolution level of $5\times$. Using this mask, large patches of size $5000 \times 5000$ pixels were processed by a number of
steps as discussed in the following sections.

![Illustration of the proposed method for cell detection and classification from a WSI](image)

**Figure 2.3:** The illustration of our proposed method for cell detection and classification from a WSI. (a) ROI detection (b) patches of size $5000 \times 5000$ pixels are extracted from ROI (c) unit which process every patch and output the coordinates and predicted label of each candidate cell (c1) cell segmentation followed by connected component analysis (c2) patch extraction (c3) classification of all the candidate cells into one of the seven classes using a trained classifier.

### 2.3.3 Cell Segmentation

To identify candidate cells from patches of size $5000 \times 5000$ pixels, we separated the cellular content from the background using the Otsu thresholding, resulting in a segmentation map for individual cells and cell clusters. By computing a threshold value independently for each patch, we observed that the cells with light colour of the cytoplasmic region were hardly fully segmented. To improve the cell segmentation, we computed a global cut-off by artificially creating a patch of size $5000 \times 5000$ pixels containing exemplary cell patches. We followed a simple process for obtaining this value which is explained in Algorithm 1. To find an optimal threshold value, the image should contain cells representing the whole population. We employed $k$-medoid clustering [42] to select exemplary cell patches from each class. We set $k = 20$ resulting in 20 clusters per class since we needed 20 exemplar patches from each class. A sample closer to the medoid of the cluster was added to the exemplar bucket. Using these exemplar patches, a big synthetic image was generated by randomly placing the exemplar patches on a plain background image retrieved randomly from one of the WSIs. This image was then converted to HSV from which the saturation channel was used to find the threshold value using Otsu thresholding. A generated segmentation map for an example visual field is shown in Figure 2.4.
Algorithm 1: Threshold selection for candidate cell segmentation

Result: threshold

1 \( k \leftarrow 20; \)
2 \( \text{exemplars\_bucket}\leftarrow[];\)
3 \( \text{synthetic\_image}\leftarrow\text{get\_random\_bg\_image}();\)
4 \( \text{for } iClass\leftarrow 1 \text{ to } \text{num\_classes} \text{ do} \)
5 \( X\leftarrow\text{rgb2gray}(\text{resize}(\text{all\_patches},128));\)
6 \( X_{\text{pca}}\leftarrow\text{pca}(\text{vectorize}(X),100);\)
7 \( \text{clusters,medoids}\leftarrow\text{kmedoid}(X_{\text{pca}},k);\)
8 \( \text{for } iCluster\leftarrow 1 \text{ to } k \text{ do} \)
9 \( \text{temp}\leftarrow\text{min\_distance}(\text{medoids}(iCluster),\text{clusters}(iCluster));\)
10 \( \text{exemplars\_bucket}\leftarrow\text{add\_to\_bucket}(\text{reshape}(\text{temp}));\)
11 \( \text{end} \)
12 \( \text{end} \)
13 \( \text{synthetic\_image}\leftarrow\)
14 \( \text{generate\_synthetic\_image}(\text{exemplars\_bucket},\text{synthetic\_image});\)
15 \( \text{synthetic\_image}_{\text{hsv}}\leftarrow\text{rgb2hsv}(\text{synthetic\_image});\)
16 \( \text{threshold}\leftarrow\text{otsu\_thresholding}(\text{synthetic\_image}_{\text{hsv}});\)

2.3.4 Cell Detection and Classification

In this study, we followed a two-step strategy for identifying different types of cells. It includes the detection of the candidate cells by thresholding, followed by the classification of the candidate cells using a trained CNN.

Training Data Preparation The annotations were obtained at the WSI level and the patches of different sizes were extracted from the images depending on the size of the candidate cells at 40× magnification. For cell clusters, the whole region surrounded by the polygon or rectangle was extracted while individual cells for which a dot was placed around the centre of the cell were captured in a different way. For a given dot, a cell segmentation mask was generated for a patch of size 500 × 500 with a dot in its centre, followed by a connected component analysis. A component having a dot inside or on its boundary was considered a candidate cell. A patch capturing the whole candidate cell was extracted and saved to the hard drive as an input to the classification network along with its label information.

Methodology Details In our approach, we applied global thresholding to segment the candidate cells, as explained in Section 2.3.3. The generated mask was further processed with hole-filling and area-based object removal to avoid artefacts. The connected component analysis was performed to compute the bounding box for each identified object in the mask. The bounding box was then used to collect input data for the classification network. For classification, we employed Xception which is the extension of inception network [43], with
Figure 2.4: Result of cell segmentation. Note that the left image is a synthetic image generated using the cell patches extracted from the WSI. The right side image shows the segmentation map (in green) overlaid on the original image.

depthwise separable convolution operations replacing inception modules. The network architecture is shown in Figure 2.5. The input image to the network was resized to a size of $256 \times 256$ pixels and was normalised by subtracting the mean from the images. We trained the network for 392 epochs with a batch size of 20 images. The network was configured by setting focal loss as a loss function and Adam function as an optimiser. The overview diagram for cell detection and classification is shown in Figure 2.3. Our code for processing a WSI of urine cytology is publicly available (https://warwick.ac.uk/fac/sci/dcs/research/tia/software/urinecyto).

Figure 2.5: The Xception architecture. Each standard convolution and separable convolution is applied with a window size of $3 \times 3$ and is followed by batch normalisation.
2.3.5 WSI-level Classification

The clinical data used in this study comprises TPS categories assigned by our cytopathologists to each WSI of a cytology slide. The ground truth (GT) risk based labels are derived from the relative risk associated with categories outlined in [19]. It is defined in relation to the extent of follow up needed which segregates the cases with a high risk of malignancy for more aggressive follow-up. We considered the stated percentage of risk to generate the GT information for the classification of samples into low and high risk cases. We put all the cases with risk of less than 50% to be in low-risk class and the cases with a risk higher than 50% to be in high-risk class. The low-risk class comprises Normal, Inflammatory, CA cases while high-risk class contains ASM and TCC cases. There were some images in our dataset that were not scanned properly and were not in focus. We excluded these images by setting a threshold on the number of all identified cells except debris in relation to the count of cells predicted as debris. Using our system, we stratified these cases with the count of atypical and malignant cells.

2.4 Comparative Method

In the proposed methodology, we performed cell detection and classification (Section 2.3.4) in two different steps. Additionally, we employed an existing approach [44] for cell detection and classification, originally proposed for object detection in natural images. We used the same pipeline as discussed in Section 2.3 except for the method (Section 2.3.4) which was replaced with this approach discussed below.

Training Data Preparation The training of an object detector requires a training set with either dense annotations or an approach to nullify the effect of unannotated objects from the loss function. In liquid-based cytology (LBC) samples, cells do not often confine to a compact region. Therefore, it results in regions with sparse annotations, not suitable for the training of object detectors. To mitigate this problem, we generated synthetic regions of dense annotations with the cells extracted from the different WSIs. First, we randomly extracted a background image of size $5000 \times 5000$ from one of the WSIs; then the cell patches used in our proposed approach were randomly placed on it. The background white patches were excluded while training this network.

Methodology Details In our second approach, we employ an object identification method for simultaneous detection and classification of cells. There are a number of one-stage and two-stage object detectors, not limited to [45], [46],...
We use a one-stage detector which has been shown to perform well in terms of both speed and accuracy, known as RetinaNet. One-stage detectors are faster than two-stage detectors but do not perform well comparatively due to the class-imbalance problem. In [44], the class imbalance problem is tackled using a novel focal loss. We used ResNet as a backbone network for our experiments. We have used a publicly available code for RetinaNet for our experiments (https://github.com/fizyr/keras-retinanet).

2.5 Experimental Results

2.5.1 Cell-level Classification

We evaluated our results obtained with Xception and RetinaNet using commonly used measures, along with results of some other CNNs i.e., VGG, MobileNet, Inception and ResNet. All these networks were initiated with the pre-trained weights for ImageNet. The RetinaNet detected more than one cell with slightly different bounding boxes against a single cell. For evaluation, we computed these measures for those predictions which lie inside the GT bounding boxes. The predicted label of a cell with the highest probability was considered a final label if there were more than one prediction against a single GT bounding box.

Among all the models, Xception outperforms all the methods on our validation set. The evaluation results of all these models are shown in Table 2.2. Figure 2.6 shows confusion matrix and ROC using Xception on our validation set. For ROC, atypia and malignant cells are considered positive classes while all other classes are considered negative classes. The area under the curve is found to be 0.99.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy</th>
<th>F1-score</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xception</td>
<td>0.834</td>
<td>0.821</td>
<td>0.820</td>
<td>0.831</td>
</tr>
<tr>
<td>VGG</td>
<td>0.825</td>
<td>0.811</td>
<td>0.809</td>
<td>0.820</td>
</tr>
<tr>
<td>MobileNet</td>
<td>0.802</td>
<td>0.783</td>
<td>0.786</td>
<td>0.803</td>
</tr>
<tr>
<td>Inception</td>
<td>0.800</td>
<td>0.779</td>
<td>0.775</td>
<td>0.798</td>
</tr>
<tr>
<td>ResNet</td>
<td>0.794</td>
<td>0.775</td>
<td>0.771</td>
<td>0.781</td>
</tr>
<tr>
<td>RetinaNet</td>
<td>0.647</td>
<td>0.609</td>
<td>0.626</td>
<td>0.642</td>
</tr>
</tbody>
</table>

Table 2.2: Comparison of different CNN based classifiers on our validation set. Note that for the methods presented in the first five rows, classification was performed using the threshold-based segmentation maps while the last row presents the evaluation results of a one-stage detector.

2.5.2 WSI-level Classification/Risk Assessment

The UMAP projection of the count of all seven categories of cells is shown in Figure 2.7. The UMAP plot shows a clear separation between high-risk and
low-risk classes. However, sub-classes have significant overlap with one another and do not separate clearly. It also demonstrates that there is variability in low-risk data points in terms of features and it is higher in comparison to samples from the high-risk class which are more closely clustered, except for some outliers. A low-risk cluster on the bottom left side of the plot shows a significant overlap between CA and inflammation which is in line with the fact that the inflammatory samples tend to show atypical features induced as a reaction to the treatment. We performed the classification of WSIs based on diagnostically important cells rather than considering all the cells. Hence we restrict to 1) the count of malignant cells and 2) the total count of atypical and malignant cells. We found the total count of atypical and malignant cells to be more discriminating as compared to the count of malignant cells only. This is demonstrated in Figure 2.8 a) and b). The area under the ROC curve with the count of atypical and malignant cells is 2\% better as compared to that obtained with the count of malignant cells.

2.5.3 Additional Experiments

As aforementioned, we used the count of atypical and malignant cells for the risk estimation. However, we also conducted some additional experiments for the evaluation of different features for the same task. We computed these features for the cells identified using the trained Xception network. We performed 5-cross validation and the split was kept the same as used in the main experiments for reporting AUC presented in Figure 2.8 a) and b). We used boosted trees as a classifier for all these experiments.
Experiment #1: Count of cells for each class  For this experiment, we considered the count of the cells as a feature set and trained a classifier. We performed this experiment with two different feature sets: feature set-1 consisted of the count of all the cells and feature set-2 consisted of the count of atypia and malignant cells only. We obtained AUC $0.816 \pm 0.049$ and $0.821 \pm 0.050$ with feature set-1 and feature set-2 respectively.

Experiment #2: Proportion of cells for each class  For this experiment, we considered the percentage of the cells as a feature set and trained a classifier. Similar to experiment #1, we performed this experiment with two different feature sets. We obtained AUC $0.822 \pm 0.071$ and $0.836 \pm 0.065$ with feature set-1 and feature set-2 respectively.

Experiment #3: Sum of cell probabilities for each class  A feature set was created by taking the sum of probabilities corresponding to each class. Since the annotations were obtained iteratively, resulting in several trained networks for cell classification. We considered the probability of recently trained two networks, resulting in 14 and 4 features per WSI in feature set-1 and feature set-2 respectively. We obtained AUC $0.810 \pm 0.076$ and $0.776 \pm 0.084$ with feature set-1 and feature set-2 respectively.

Experiment #4: Phenotyping of atypical and malignant cells  In this experiment, we conducted phenotyping of atypical and malignant cells. To this end, we extracted local binary features of the atypical and malignant
Figure 2.8: Demonstrating the significance of the count of atypical and malignant cells for binary classification (low-risk vs high-risk) at WSI level. a) ROC curves using a count of malignant cells (M) and b) ROC curves using a total count of atypical and malignant cells (A+M). ROC is shown with 5-cross validation and the average area under curve using M and A+M is 0.81 and 0.83 respectively. c) A violin plot displaying the count of predicted atypical and malignant cells in WSIs belonging to low and high-risk classes. To generate these results, cells were classified using the Xception network.
cells followed by clustering. We performed cell segmentation before feature extraction and employed k-means clustering analysis for grouping the cells with similar texture representations. We experimented with a range of values for k from 3 to 15. We computed the centroids of the clusters from two different sets of cells: 1) manually annotated cells and 2) network (Xception) predicted cells. For computing centroids from the network predicted cells, we randomly selected 26 WSIs labelled as CA, ASM and TCC with the maximum number of predicted atypical and malignant cells. We used the Euclidean distance between the centroid and the feature vector to assign each cell its cluster. For WSI classification, we considered the proportion of cells in each cluster (for k 3 to 15) to train a classifier. We achieved good performance for k = 7 with AUC 0.84 ± 0.053 and 0.820 ± 0.058 obtained with centroid computed from manually annotated and network annotated cells respectively.

2.6 Discussion

The automatic cell classification in urine cytology images is a challenging task due to various reasons. This may include inconsistent annotations, classes with a subset of similar features, how the cell samples are categorised into different classes and changes in the cell appearances due to the treatment. The confusion matrix as shown in Figure 2.6 demonstrates that the atypical and malignant cells are mostly confused with each other and this is due to their overlapping visual features with respect to TPS criteria. Most of the misclassified cells belong to the ‘Others’ class and this is mainly due to the nature of the samples, we have placed in this class. It comprises degenerating cells however these cells could belong to a normal, atypical or malignant class. Normal cells change their appearance when the patient is on medication and are termed reactive cells and may resemble atypical cells. Therefore, there are some normal cells misclassified as atypical cells. We also verified the network predictions for some benign cases for which the number of malignant cells predicted by the network was greater than 10. We found that some of these cases had reactive normal cells and cells with fluffy cytoplasm. This could be improved by adding these challenging cases to the training set.

In this study, we demonstrated the potential promise of an automated risk stratification method. There are some limitations of the proposed method related to how the data was obtained. We obtained WSIs and their corresponding data from a single centre which may have introduced a bias in our proposed approach. Therefore, our findings need to be validated with a large-scale study. Additionally, sourcing annotations from two pathologists may also have introduced a bias into the machine learning model. However, involving more pathologists is not necessarily the solution to this problem due to the
Table 2.3: Table presenting intra-observer variability of cell-level annotations provided in two attempts (Att# 1 and 2) by a pathologist. The Cohen’s Kappa is 0.15 showing a slight agreement for these classes; mainly the disagreement is in between atypical and malignant classes and in between atypical and others classes.

potentially larger degree of disagreement between multiple pathologists, as shown in previous studies such as Reid et al. [52].

2.6.1 Annotation Variability

We sourced cell-level annotations from two pathologists. The inconsistency in their annotations can undermine the performance of the model, given the model tends to learn the complexity. To inquire about the inconsistency in the labelled dataset, we randomly selected some cells from our validation set and asked the expert pathologist to re-annotate them. We selected these cells from our more concerned classes, normal, atypia and malignant. We selected atypia and malignant classes since these are important in terms of making a diagnosis. The normal class was selected since it was mostly misclassified as atypia by the network. The variability in the annotations of the same pathologist is demonstrated in Table 2.3. In addition to the slide quality and the lack of multiple focal planes, the intra-observer variability could be due to pathologists’ lack of experience with the digital slides for urine cytology. The intra-observer variability is a recognised issue in cytology. However, sourcing the labelling with consensus among different pathologists to improve the variability will improve the performance of the model.

2.6.2 Performance of RetinaNet

There is a huge difference between the performance of RetinaNet with ResNet and a ResNet followed by the cell segmentation. This is partially due to the limitation of the detector in the RetinaNet, missing several cells. Additionally, the detector resulted in many bounding boxes for a single candidate object. Choosing a bounding box with a probability greater than a certain threshold value further increases the number of missing cells. In our validation set, we had 5175 cell samples, out of which 68 cells were missed when no detected object was ignored. However, selecting the predictions with a probability greater than 50% resulted in 692 cells being missed. Contrary to it, the threshold-based
segmentation does not miss any cell, except that it may fail to segment the whole cell, particularly the squamous cell.

### 2.6.3 Correlation between Cytology and Histology

We also studied the correlation between the cytopathology based diagnosis and histopathology based diagnosis. We obtained histopathology diagnoses for 48 cases along with their cytology and histology reports. These cases comprise 26 CA and 37 ASM, diagnosed using the cytology slides. We hypothesised that the cases for which the network predicted more atypical and malignant cells would be diagnosed as malignant on performing histology. We observed a trend of association between cell count and histopathology based diagnosis as shown in Figure 2.9. We compared the results of the cytopathology based risk category and our digital risk labelling against the ‘gold standard’ histopathology based diagnosis. The confusion matrix for manual cytopathology based risk vs manual histopathology based diagnosis is shown in Table 2.4. As can be seen in Figure 2.10, the digital risk could be considered a better predictor of the histopathology based diagnosis. However, this needs to be validated with a large-scale multi-centre study. To study it further, we looked into the cytology and histology reports of some of these cases to understand the grounds for the possible discrepancies between cytology and histology diagnosis. We came up with the following rationales for the discrepancies: 1) Urine of patients with bladder cancer can be negative (i.e. no shedding of malignant cells in urine). 2) In cases of instrumented urine samples, at least some abnormal looking groups of cells can be expected due to this sampling technique. Also, an intervention or surgical procedure can lead to the appearance of granulation tissue, inflammation and reactive atypia. However, if information about the sampling technique and relevant history is not available to a pathologist, these cells can then get wrongly labelled as atypical or suspicious. 3) Information about female genital tract, kidney or prostate pathology is relevant and should be available to pathologists. Otherwise malignant cells from these organs (which can sometimes be found in urine) could be misdiagnosed as malignant urothelial carcinoma cells (i.e. bladder carcinoma). 4) The presence of calculi/stones or the BCG treatment also results in the appearance of abnormal-looking cells, thus affecting the cytology diagnosis. 5) Malignant cells can be missed in specimens contaminated by fungal or bacterial overgrowth. 6) In histology reports, both high and low-grade tumours are reported while in cytology detecting low-grade TCC is not appreciated. 7) A long interval between cytology and histology can be one of the reasons for the difference in diagnosis.

32
Figure 2.9: Scatter plot demonstrating the relation between the histology-based (Benign and Malignant) and cytology-based (CA and ASM) diagnosis. Each marker represents a case. The urothelial cells were classified using the Xception network.

<table>
<thead>
<tr>
<th>Cytopathology based Diagnosis</th>
<th>Benign</th>
<th>Malignant</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>10</td>
<td>16</td>
</tr>
<tr>
<td>ASM</td>
<td>8</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 2.4: Confusion matrix for manual cytopathology based risk vs manual histopathology based diagnosis (Sensitivity = 0.47, Specificity = 0.56).

2.6.4 Inter-observer variability w.r.t. WSI labelling

All pathologists involved in labelling are fellows of the Royal College of Pathologist (RCPath) with experience in assessing urine cytology slides under a light microscope. However, it does not suggest a lack of disagreement in their diagnosis. Keeping this in mind, we investigated the concordance of our digital risk with the diagnosis of an independent pathologist. We selected WSIs which were misclassified (low-risk and high-risk cases with high and less number of atypical and malignant cells respectively) by our method w.r.t. the reference labelling. We selected 20 such cases from low-risk (CA) and high-risk (ASM) categories and asked the independent pathologist (referred as Pathologist-A) to assign labels to these WSIs. Figure 2.11 shows the inter-observer variability between the reference and an independent pathologist-A labelling. Out of 20 cases, the agreement was found for only 5 cases. In other words, our predictions make a concordance of 0% and 75% with reference and pathologist-A respectively. The readers are referred to Table 2.5 for more details on inter-observer variability. This disagreement could be because the reference labelling was carried out in a clinical setting where pathologists had access to other clinical data whereas the pathologist P-2 made their decision solely based on image
content. Also, voided and aspirated slides are interpreted slightly differently by the pathologists. P-1 pathologists knew how the sample was obtained while the pathologist P-2 did not have this information and interpreted all the samples as if they were voided samples. We observed inter-rater variability (as shown in Table 2.5) more between cytological atypia and suspicious cases, both of which are already considered contentious and borderline, rather than between malignant and normal cells. This is similar to findings reported in [25–27]. Considering this variation, the ROC obtained in this study could vary on testing the proposed method with WSI labels obtained from a different pathologist. The inter-observer variation in labelling cells and WSIs makes the automated diagnosis of cytology samples challenging.

2.7 Summary

In this study, we found that the count of atypical and malignant cells is more robust in discriminating between low and high-risk cases as compared
Figure 2.11: A Venn diagram presenting inter-observer variability in labelling the 20 WSIs. These cases were misclassified by the method w.r.t. the reference but 15 out of those predictions agreed with another pathologist-A. The term reference represents the diagnostic information obtained from the UHCW while pathologist-A is an independent pathologist from a different hospital. To generate these results, cells were classified using the Xception network.

Table 2.5: Table presenting inter-observer variability in labelling the WSIs. Considering the two class classification (low and high risk), the agreement was found on 5 cases only out of 20 cases between P-1 (Reference) and P-2 (Pathologist-A).

to the count of malignant cells only. The difference between the clinical study and our finding is due to several interdependent factors including intra-observer variability in annotations for atypical and malignant cells, leading to poor performance of the classifier in discriminating between the atypical and malignant cells. Since the cytology material is less evenly distributed (even in LBC samples), pathologists frequently need to focus on different planes to view all the cells in a cell cluster. Therefore, due to the intrinsic nature of cytology samples, the z-stacking feature can potentially help. We believe that the availability of different planes in the images could improve the annotations and network performance at the same time. The proposed method for automated risk stratification of urine cytology slides has demonstrated clear promise. However, before we can deploy such a system in clinical practice, we will need to conduct large-scale multi-centric trials for establishing the efficacy of the proposed method.
Chapter 3

Handcrafted Feature
Matching based Registration

3.1 Introduction

Registration often serves as an essential pre-processing step for many medical image analysis tasks. A typical approach for registering two images consists of an optimisation algorithm accompanied by a similarity measure, selected based on the image capture modality. The optimisation algorithm finds the best spatial transformation by maximising the similarity measure which evaluates the correspondence between the images after applying the transformation. In digital pathology, registration has typically been used to capture information from a single modality images IHC stained using different biomarkers and often also including the routine H&E slides too. Broadly speaking, histology image registration has three main applications: cross slide image analysis [53, 54], multi-modal image fusion [55] and 3D reconstruction of serial sections [56–59]. In this thesis, emphasis has been given to the registration of a stack of consecutive multi-stain histology images for multi-slide image analysis.

To proceed with the multi-slide analysis, cross-slide alignment of the WSIs for serial sections is a prerequisite. This is because, during the slide preparation process, tissue sections cut from the same tissue block will not retain their continuity in the z-axis, an example shown in Figure 1.2. Therefore, registering these images is an important step before any automated multi-slide colocalisation analysis [53, 54, 60]. The cross-slide registration of histology images is a challenging task due to many reasons: changing structure between the sections, missing parts, tissue folds, broken tissue and even the overall morphology of the tissue could change due to their fragility. There is a need for a registration approach that can perform well under these conditions and can align pairs of images in a reasonable time to facilitate the downstream analytical or diagnostic pipeline.
Cross-section slides mostly have non-linear deformations that cannot be tackled with a global transformation alone. Whereas, non-rigid registration methods are capable to find correspondence by locally transforming a source image. Although there is a wide range of registration methods proposed in the literature, their uptake in the commercial software is lagging due to the lack of robustness [61]. Non-rigid registration with Newton-type optimisation relies on an initial guess that is close to the optimum to obtain a fast convergence rate and also to prevent converging to local minima. Therefore, a good choice of global linear transform is required as an initialiser to unify its accuracy with the robustness of the non-rigid transformation. Although, there are studies suggesting pre-registration methods using manual alignment [62–64], an automatic approach for linear alignment without any user input is recommended. In this chapter and the next chapter, we present two different registration methods for automatic estimation of global linear transformation between the two images.

In general, there are two main methods for automatic linear image registration: intensity-based registration and feature-based registration [65–67]. As the name suggests, intensity-based techniques directly utilise the pixel intensity information in an image pair and optimise a similarity metric to find the correspondence. The correspondence is found by transforming a moving image such that it maximises the similarity measure between the reference image and the transformed moving image. Feature-based methods would first identify the key features from the images and then a transformation is estimated by a matching system by utilising the matching features between two images. The choice between the two methods depends on the nature of the images. However, feature-based methods are often considered better in handling geometric transformations [68]. In this thesis, we are analysing slide images stained with different biomarkers which means that the staining pattern would vary among them. The intensity between differently stained IHC slides can be non-linear and it would be likely that we will not observe any correlation between intensity values of the same tissue region stained with two different biomarkers [60]. Therefore, intensity-based methods generally struggle to compute accurate transformations [65]. Moreover, these methods also suffer in the presence of other artefacts, missing tissue regions or tissue tearing [2, 57]. For this reason, a feature-based approach is considered more suitable for cross-slide registration.

There is a significant amount of work done with handcrafted features for an alignment task [2, 54, 62, 69–73]. Such approaches generally consist of four steps: 1) detection of salient features, 2) calculating descriptors from pixels around the detected features, 3) finding matching descriptors between the two images and 4) finding the transformation, mapping the matched features of a moving image to the corresponding features of a reference image [74].
The Scale Invariant Feature Transform (SIFT) [75] has been widely used for pathological images [59, 62, 62, 73, 76–78]. It is also incorporated in ImageJ plugin for registration to extend the bUnwarpJ method [79].

In this chapter, we discuss our work with handcrafted feature matching for the registration of multi-IHC images. We extend an existing method on handcrafted feature matching using maximally stable extremal regions (MSER) [80] which has been successfully adapted for multi-stain histology images [2] and for which the implementation code was readily available. We employed this method in a multi-scale manner; at first registering images at a low resolution by matching MSER features between an image pair, followed by a refinement step performed at the immediate higher magnification level, using the same feature matching technique. Our results demonstrate that the multi-scale approach can improve the overlap between reference and moving images by estimating a good alignment.

3.2 Literature Review

The registration of histology images is generally a challenging problem due to several aforementioned reasons and should be completed in a reasonable amount of time. In literature, there are several methods proposed to tackle both linear and non-linear deformation in the tissue. In [81], Mueller et al. demonstrated the feasibility of non-rigid spline-based registration for WSIs. The authors employed it for the alignment of serial tissue section WSIs with different staining characteristics such as tissue sections stained with IHC biomarkers and H&E stains. Their proposed approach follows a two-step multiscale strategy to estimate the transformation within a reasonable time constraint. In the first stage, the initial transformation map is estimated using a publicly available tool for registration, known as Elastix. Then in the second stage, the initial transformation is applied to the high-resolution regions of WSI rather than the whole WSI. Schwier et al. [69] proposed a rigid registration method based on finding the best matching features of the vessel structures between a pair of images.

Edge-based registration methods are generally computationally less expensive than registration methods based on pixel intensities. Hierarchical Chamfer matching and registration based on curvature scale space (CSS) representation of the boundary points are common examples of edge-based registration. These methods are fast since the transformation is estimated based on boundary points rather than pixel intensities. In [71], Trahearne et al. employed curvature scale space (CSS) representation of tissue boundaries for WSI registration. The existing CSS representation is adapted by replacing zero crossings with the curvature maxima. However, this approach may not be reliable when
one of the images in a pair has a broken or missing tissue part. In addition, this approach requires optimal tissue segmentation and for bigger tissues, it is harder to segment the tissue boundary due to overlapping tissue with a cover slip hence it is unlikely to perform better. In another study [54], the authors used CSS based method for pre-alignment of serial sections stained with different biomarkers. The results of pre-alignment are further improved based on nuclear clusters and fatty regions since these tissue structures are more likely to exist across several serial sections. In [2], a novel approach based on MSER was proposed and employed for the alignment of multi-IHC CRC sections. The proposed approach extracts features for each MSER detected from the preprocessed Haematoxylin (H) stain channel. These features are utilised for finding the best pair of MSERs and are used as corresponding control points for estimating rigid transformation. The authors employed this approach for pairwise registration of consecutive sections and also for case-wide registration for finding the optimal order of the sections for estimating the transformation of the whole stack. In case-wide registration, pairwise transformations are first computed for all the distinct pairs of sections, which are then evaluated by determining the transformation consistency, based on the transitive property of rigid transformations. Using the graph representation of slides and consistency matrix, the best sequence of pairwise transformations is selected by computing the maximum spanning tree.

Obando et al. [82] proposed a method for quantifying T cells which could help develop immunotherapy. Registration is performed to align two images: one stained with cytokeratin to mark the tumour cells and the other stained with CD3 to label T cells. The registration is carried out in three steps: extraction of common information using Kmeans, preliminary rigid registration followed by B-spline registration on down-sampled images and application of B-spline deformation model to full-size images. The application of this work is limited due to two reasons: selection of a threshold for step 1 is manual and it is assumed that there are no tissue folds or tissue losses. Solorzano et al. [62] proposed a semi-automatic piece-wise strategy for registering histology images using rigid transformation. Corresponding sub-regions/pieces from the two tissue sections are manually selected using a web interface. These pairs of sub-regions are then aligned individually using SIFT and Multi-scale Oriented Patches (MOPs) feature matching approach. Borovec et al. [72] evaluated eleven existing registration methods for both linear and non-linear transformations on a benchmark dataset comprising 616 image pairs.

More recently, work on registering histology images has gained momentum after the organisation of the Automatic Non-rigid Histological Image Registration (ANHIR) challenge aiming at the non-linear registration of multi-stain microscopic images. This challenge was organised in conjunction with IEEE
ISBI 2019 conference [1]. In total, ten teams submitted their results for the challenge and all these submissions were made using the existing methods. In [1], authors have provided evaluation of seven best-performing methods, along with six well-established methods (bUnwarpJ, RVSS, NiftyReg, Elastix [83], ANTs and DROP). All seven teams performed registration using greyscale images in two stages. First, pre-alignment was performed to estimate a linear transformation either using the intensity-based approach or feature-matching method, followed by a non-linear transformation. The winning team (MEVIS) followed a three-step approach: pre-alignment, parametric registration and non-parametric registration [84, 85]. They employed a Normalised Gradient Fields (NGF) distance measure [86]. The transformation estimated at any step is used as an initialiser for the subsequent step. In the pre-alignment step, translation offset is estimated by taking a difference between the centre of mass of both images. The angular offset is computed by trying several rotations and the one with the minimum NGF measure is selected. In the second step, the Gauss-Newton method is used for computing an affine transformation while in the third step, a dense non-linear transformation is estimated using L-BFGS optimisation. A summary of methods used by all seven teams is shown in Table 3.1. Only one team (TUB) used CNN based approach for registration and performed well on the training set rather than on the evaluation set due to limited generalisability. It also appeared to be the fastest performing method, better than the winner in terms of average time.

![Table 3.1](image)

Table 3.1: A summary of methods used in the ANHIR challenge by the seven participating teams; NGF (normalised gradient field), NCC (normalised correlation coefficient), MI (mutual information), L-BFGS (limited memory Broyden-Fletcher-Globfarg-Shanno).

### 3.3 Hand-Crafted Feature-Based Registration

In this chapter, we present a hand-crafted feature-based method for the registration of multi-IHC histology images. We experiment with MSER features [80] which have been shown to perform well for multi-stain WSIs with some limitations [2]. First, salient features are extracted from a pair of images,
followed by finding the matching features between the two. We assume that the spatial locations of matched features correspond to the same region in the pair of images. These matching features’ locations are used as control points for estimating the geometric transformation between the two images. We extend the existing approach of MSER feature-based registration [2] in a multi-scale fashion and show that it helps reduce the registration error.

### 3.3.1 Maximally Stable Extremal Regions

MSER is a feature detector aimed at finding stable connected components by applying multiple thresholds on a greyscale image. If we threshold a greyscale image with increasing threshold values, then we expect to see binary structures growing in size. MSERs are structures that do not vary significantly under threshold changes and are considered stable. There are a number of parameters that can be adjusted to control the stability of MSERs and would influence the number of MSERs detected. These parameters are:

- Threshold step size ($\Delta$) defines an increment in threshold values. Its typical value ranges between 0.8 and 4; decreasing this value returns a high number of MSERs;

- Stability score ($\omega$) defined by a maximum change allowed in the number of pixels under threshold changes. Its typical value ranges between 0.1 and 1.0 and decreasing this value would return a smaller number of MSERs. However, the size of these regions is likely to persist with varying thresholds, hence these regions would be considered more stable than one would get with a higher stability score;

- Minimum number of pixels ($a_{\text{min}}$): this parameter will allow the selection of MSERs with a number of pixels greater than this value;

- Maximum number of pixels ($a_{\text{max}}$): this parameter will include only those regions with a number of pixels less than this value.

### WSI Registration Using MSER Features

In histology images, MSER can represent nuclei, glands or bigger structures depending on the magnification level. MSER detection for a WSI is not practical in terms of computational time and limited memory resources. Therefore, it is best to apply it to a downscaled version of WSI which can be loaded into memory. In [2], MSER was applied to images downscaled by the factor of $2^7$. The MSER algorithm was applied to the H stain channel after histogram equalisation as a preprocessing step. The parametric values used for MSER algorithm in [2] are: $\Delta = 1.5$, $\omega = 25\%$, $a_{\text{min}} = 125$ pixels and $a_{\text{max}} = 10,000$. 

41
pixels. There are several steps for matching corresponding MSERs, as shown in Figure 3.1. These steps are discussed below:

1. Feature Extraction: For each MSER, four features were computed: number of pixels, \( x \) and \( y \) coordinates of the region centroid and orientation of the best-fit ellipse enclosing the region.

2. Selection of Valid Pairs: The pairs of MSERs, one from the reference image and the other from the moving image are considered valid pairs if their area is similar. These valid pairs are likely to correspond to the same tissue component that can be used for estimating candidate transformation. There are chances that MSER valid pairs with similar properties may belong to a different tissue region. Therefore, all the valid pairs are tested to check their validity for registration.

3. Estimating Candidate Transformation: The first step in checking the validity of registration is by computing a transformation using the features of a corresponding valid MSER pair. The vector pointing from the centroid of the moving MSER to that of the reference MSER is used to define a translation offset. The rotation angle is determined by taking the
difference between the orientation of the MSER valid pair. Since elliptical orientation is undirected, therefore the rotational offset is calculated in both directions. This would generate two candidate transformations for each valid pair.

4. Selection of Aligned Pairs: The candidate transformation results in new pairs of MSERs and the best transformation is the one generating MSER pairs with a minimum angular difference. In [2], MSER pairs generated after transformation are considered aligned if their angular difference is no more than $15^\circ$.

5. Selection of Matched Pairs: It is not uncommon for MSER pairs to correspond to the different tissue regions and yet have similar areas and orientations. Therefore, there is a need to find the best subset of aligned pairs ($\beta$) describing the same tissue region. This subset is referred to as matched pairs in [2]. Additionally, the matched pairs should spread all over the tissue section since the quality of registration would be adversely affected if the matched pairs are confined to one part of the section. To this end, in [2], all the aligned pairs are first sorted by the Euclidean distance between the centroid of MSERs. Then, an aligned pair is added to the matched set if the following two conditions are met: neither MSER of the pair exists in the list of matched pairs with a shorter distance ($\bar{R}$) and the pair is not in close proximity to others in the matched pairs. In total, 20 MSER pairs ($\beta = 20$) are included in the list of matched pairs, with minimum distance among them to be 50 pixels ($\bar{R} = 50$) [2].

6. Quality Check: Steps 3-5 are performed against each valid pair since each valid pair is a candidate for computing the final transformation. Registration candidates with less than 20 matched pairs are excluded and for the rest of the registration candidates, the quality is determined by taking a sum of the distances between the matched pairs. The matched pairs with less distance are likely to perform a better alignment.

7. Transformation Estimation: The centroids of the matched MSER pairs, corresponding to the best registration candidate, are used to generate a set of corresponding points. These points are used as control points for estimating the transformation parameters using a least squared approach.

### 3.3.2 Multi-scale MSER

In the previous section, we discussed the application of MSER features for registering WSIs at a single scale ($S = 2^7$). Here, we propose a multi-scale version of MSER matching based registration. We show that the multi-scale
based approach leads to a more robust and accurate alignment as compared to the single-scale based method.

3.3.3 Pre-processing steps

We utilised the same pre-processing steps as were employed by the single-scale MSER based registration, proposed in [2]. Entropy filtering was performed on the greyscale image followed by Otsu thresholding for generating initial tissue segmentation masks. To improve these initial masks, local binary pattern features were used. To this end, initial tissue segmentation was divided into 1000 × 1000 blocks. LBP map and its corresponding histogram were computed. SVM was then trained using these block histograms to distinguish the non-tissue area from the tissue region. In our experiments, we generated tissue masks using a simple UNET shaped architecture without any skip connections.

Stain Deconvolution

Similar to the previous work [2], we have used the original MSER algorithm compatible with greyscale images. To acquire correct matching MSER points between two images, the intensity pattern should be similar such that similar MSER features are detected. Although RGB images express more useful information as compared to their greyscale version; the multi-stain images used in our work contain different staining patterns and hence the application of the MSER algorithm may result in incorrect matching MSER points. Therefore, to acquire input images with consistent intensity patterns across the stack of images, the H stain channel is a suitable candidate as it can be found in each slide irrespective of the additional IHC biomarker. We performed a stain deconvolution operation with a fixed stain matrix for the extraction of the H channel.

Tissue Segmentation

We used a simple CNN architecture for tissue segmentation, as shown in Figure 3.2 (c). Each convolution layer is followed by the ReLU activation function. We created a ground truth tissue mask for 65 WSIs downsampled to magnification level 1.25×. We used 62 images for training and 3 for validation. These images were stained with H&E and different IHC biomarkers. For consistency, we extracted H stain from all these images rather than training the model with the original RGB images. Instead of creating a binary mask representing tissue as foreground and non-tissue area as background, we considered having three classes. These three classes correspond to a tissue region, glass slide (mostly white) and slide artefacts (slide label and glass slide artefacts). An example image with its mask is shown in Figure 3.2 (a,b). In this thesis, we will refer
to this tissue mask as **TS_1**. Adam optimiser was used to train this model using input patches of size 512 × 512 pixels at magnification level 1.25× with a batch size of 4. Mean square error was used as a loss function. Similar to [2], the control tissue was removed from all the images.

![Image](image_url)

**Figure 3.2**: The input images and architecture of the CNN model used for tissue segmentation. a) H stain image downsampled at magnification level 5, b) ground truth tissue mask and c) CNN model. The colour blue, green and red shown in b) represent background, slide artefacts and tissue region, respectively.

### Histogram Matching

Histogram matching [92] was performed to deal with differences in intensities and contrast between the reference and moving images. It is a lightweight normalisation technique that calibrates a moving image such that its histogram matches that of the reference image. In the process of normalisation, the non-tissue area (glass slide, slide label and other artefacts) negatively affects the output if they are not consistent between the image pair. A tissue mask was used to mask out such areas before constructing a histogram. To perform histogram matching for a pair of images, we need to select one of the images as a reference and the other as a moving image. We made this selection based on the variance of tissue pixels. An image with high contrast appears sharp and tends to have a high variance. Therefore, an image with high variance was considered a reference image. After histogram matching, the contrast of both the images was further improved by linearly rescaling the values while saturating the lower 1% and top 1% of all the pixel intensity values. In other words, the lower 1% of the pixel intensities values are converted to 0 while the top 1% of them become 255.
Figure 3.3: An example of input image preprocessing steps. The variance of the reference H channel is greater than that of the moving H channel. Therefore, during the normalisation process, the H channel of the moving image is modified such that its histogram is similar to that of the reference H channel. A tissue mask was used to exclude the non-tissue area while constructing the histogram.

Registration using Multi-scale MSER Features

We extended the previously proposed MSER based registration approach [2] in a multi-scale manner. First, we apply this method to a WSI at a magnification level of 0.3125×. This is referred to as ‘global transform’. Then, we extract the tissue region from the reference image and the transformed moving image at a magnification level 0.625× and perform steps 1-7 as in Section 3.3.1 to estimate the best transformation for the moving tissue region. This is referred to as ‘tissue transform’. Next, we extract the transformed tissue region from the moving image and the corresponding tissue region from the reference image at a magnification level 1.25×. At this stage, the processing time would increase significantly due to the size of the image pair. To expedite it, we follow the divide and conquer approach by dividing the tissue region into four equal parts. We perform steps 1-7 for each tissue part separately in a parallel fashion. This would generate the best matching pair for each tissue part. The centroid of all the matched MSER pairs is then considered as control points for estimating the final transformation parameters using a least square approach. This is referred to as the ‘block-wise’ transform. We used the same parametric values, as mentioned above, except for $\Delta$ and $\bar{R}$. We automatically select a value of $\Delta$ for each case such that the number of MSERs is less than 1500. This is to
ensure that the matching process is completed in a reasonable time. The value \( R \) varies along the different magnification levels; 50, 100 and 1000 pixels at 0.3125\( \times \), 0.625\( \times \) and 1.25\( \times \) magnification levels, respectively.

### 3.4 Datasets & Performance Measures

We evaluated our multi-scale approach on a subset of the COMET dataset. A detailed explanation of the dataset used for the registration task and the related evaluation metrics is presented in the following subsections.

#### 3.4.1 Datasets

COMET dataset is obtained from the University Hospitals Coventry and Warwickshire (UHCW) NHS Trust in Coventry, UK. This dataset comprises WSIs of 86 cases, taken from different patients. There are 16 slides per case, each scanned using the Omnyx VL120 scanner at 0.275 microns/pixel. These slides are stained with different stains: CK8/18, Ki67, p53, Vimentin, MMR (MLH1, MSH2, MSH6 and PMS2), Ecadherin, EpCAM, PTEN and H&E and their exact sequence of staining is shown in Figure 3.4. We selected a set of 7 cases and present the quantitative evaluation of registration methods using them. We considered six slides per case, involving MMR prediction from H&E and CK8/18 images for the end purpose of MSI prediction. Therefore, we performed registration of MMR slides with CK8/18 and H&E slides. We also aligned H&E slide images to CK8/18 slides to generate a corresponding tumour mask for CK8/18. Among various difficulties confronted while registering tissue sections, the extent of differences in tissue structure makes WSI registration a challenge. These differences vary between sections depending on how far they are cut from the tissue block relative to the target section. For the same reason, the alignment of H&E w.r.t. CK8/18 was challenging as there were 10 sections in between them, including the MMR markers’ sections. While CK8/18 and MMR biomarker slides are the consecutive sections sliced in a given order: CK8/18, MLH1, MSH2, MSH6 and PMS2. These IHC stained slides tend to be highly correlated in terms of tissue structures which is not the case with H&E since it is around 50\( \mu \)m away from the given IHC slides. Therefore, significant variability in the tissue structures exists among the H&E and IHC stained slides, as shown in Figure 3.5. For evaluation, we considered 15 pairs of sections per case for registration: aligning MMR markers w.r.t CK8/18 (4 pairs), aligning all IHC slides w.r.t H&E (5 pairs) and all possible combinations of MMR biomarker slides (6 pairs). This resulted in 105 pairs in total against 7 cases.
Figure 3.4: An example to illustrate the exact sequence of sections cut from a tissue block in the COMET dataset.

Figure 3.5: Similarities and dissimilarities in the tissue architecture among spatially corresponding visual fields extracted from a) CK8/18, b) MLH1 and c) H&E stained images. a and b are taken from the consecutive slides (5µm apart), hence has more similarities as compared to a and c or b and c (more than 50µm apart).

3.4.2 Performance Measures

To evaluate the performance of registration methods, we manually defined landmarks on all the images selected for evaluating registration performance. We used the ImageJ tool for annotating significant tissue structures with landmarks. A landmark was marked if the corresponding tissue was found in all the images. Figure 3.6 shows landmarks overlaid on patches extracted from the corresponding slides. We defined two sets of landmarks, referred to as LandmarkSet-1 and LandmarkSet-2 in this thesis. LandmarkSet-1 comprises a set of 50 landmarks for each slide and is used for the registration of IHC slides, whereas LandmarkSet-2 comprises a set of 10 landmarks per slide for the registration of H&E images to IHC images. The reason for having two sets of landmarks is that it may be unlikely to observe similar tissue structures among IHC and H&E images. To evaluate the quality of registration, we compute the target registration error (TRE) for image pairs in the test set \( T \). For an
image pair $j$, we compute the distance-based error measure using the following formula

$$TRE(R_j, M'_j) = ||R_{\text{land},j} - M'_{\text{land},j}||_2$$

(3.1)

where $j \in \mathcal{T}$. $R$ and $M'$ belong to an image pair $j$ and represent reference and transformed moving images, respectively. $R_{\text{land}}$ and $M'_{\text{land}}$ denote landmarks of reference and transformed moving images, respectively. $M'_{\text{land}}$ is obtained by applying the estimated transformation on landmarks of moving image $M_{\text{land}}$. The registration error is normalised by the length of the reference image diagonal so that the amount of error is independent of the image size.

$$rTRE(R_j, M'_j) = \frac{TRE(R_j, M'_j)}{\text{hypot}(R)}$$

(3.2)

where $\text{hypot}(R) = \sqrt{w^2 + h^2}$ and $w$ and $h$ denote width and height of the reference image, respectively. The above equation for computing $rTRE$ generates a list of values for an image pair $j$ which we aggregated by taking their median. Overall registration error for $\mathcal{T}$ was computed by either taking mean or median of the aggregated $rTRE$, namely as the median of median $rTRE$ (MMrTRE) and the average of median $rTRE$ (AMrTRE). We also report the average of the maximum $rTRE$ (AMaxrTRE).

$$MMrTRE = \text{median}\left\{\text{median}_{j \in \mathcal{T}}\{rTRE(R_j, M'_j)\}\right\} \quad \forall \ j \in \mathcal{T}$$

(3.3)

$$AMrTRE = \text{mean}\left\{\text{median}_{j \in \mathcal{T}}\{rTRE(R_j, M'_j)\}\right\} \quad \forall \ j \in \mathcal{T}$$

(3.4)

$$AMaxrTRE = \text{mean}\left\{\max_{j \in \mathcal{T}}\{rTRE(R_j, M'_j)\}\right\} \quad \forall \ j \in \mathcal{T}$$

(3.5)

We also tested the robustness of registration results by comparing the transformed landmarks with the initial landmarks before any alignment. We computed the robustness in a similar way defined by the ANHIR challenge organisers which is the relative number of successfully transformed landmarks. Any given landmark pair $R_{\text{land},i}$ and $M'_{\text{land},i}$ is considered to have been registered successfully only if the distance between them is smaller than the difference between $R_{\text{land},i}$ and $M_{\text{land},i}$. Robustness for an image is computed by counting the number of successfully registered landmarks divided by the total number of landmarks for that image. Robustness over the whole dataset is computed by taking the mean over all the image pairs’ robustness.
3.5 Experimental Results

3.5.1 Experimental Setup

For single-scale and case-wide approach, we have used the same parametric values as in [2], $\Delta = 1.5$, $\omega = 25\%$, $a^{\text{min}} = 125$ pixels, $a^{\text{max}} = 10,000$ pixels, $\beta = 20$ and $\bar{R} = 50$. For our multi-scale approach, we have modified the values for $\Delta$ and $\bar{R}$, as discussed in Section 3.3.3.

3.5.2 Results

Table 3.2 shows quantitative results after each step in our multi-scale approach, along with the initial registration error obtained before any transformation. These results demonstrate that the alignment is refined by performing the registration on more than one scale. The registration error is reduced for all three evaluation measures. In our experiments, $AMrTRE$ is significantly reduced when feature matching is performed using tissue regions cropped at magnification level $0.625 \times$. The improvement can be attributed to the detection of more MSERs as we increase the resolution and hence higher chances of getting better matching pairs. We have also shown these step-wise improvements for each case in Figure 3.7. For almost all the cases, refining registration at an immediate higher resolution (Tissue Transform) improved the registration accuracy. However, transformation computed using a block-wise step didn’t make a significant improvement. Also, block-wise transformation cannot improve alignment if previous transformations were unsuccessful and it is often due to the lack of correspondence between the pairs of tissue parts.

We also tested the performance of our multi-scale approach for registration against two methods: 1) single-scale approach and 2) case-wide approach. Table 3.3 gives the registration results for IHC-stained image pairs and IHC vs H&E image pairs separately. This is to show how these methods perform for the challenging cases with significant changes in tissue structure between two images (IHC vs H&E). For IHC vs IHC, the multi-scale approach outperforms
Table 3.2: Demonstration of improvement in registration error on applying feature matching at different resolutions in a multi-scale approach. The feature matching is applied for a WSI downsampled to a resolution level 7 for global transform. While for tissue and block-wise transforms, it is performed for the tissue region only at a resolution level of 6 and 5, respectively. Due to its large size, the tissue region is divided into four blocks at resolution 5.

<table>
<thead>
<tr>
<th></th>
<th>Avg(Med(rTRE))</th>
<th>Med(Med(rTRE))</th>
<th>Avg(Max(rTRE))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>0.1797</td>
<td>0.1068</td>
<td>0.2323</td>
</tr>
<tr>
<td>Global Transform</td>
<td>0.0162</td>
<td>0.0028</td>
<td>0.0312</td>
</tr>
<tr>
<td>Tissue Transform</td>
<td>0.0072</td>
<td>0.0024</td>
<td>0.0310</td>
</tr>
<tr>
<td>Block-wise Transform</td>
<td>0.0068</td>
<td>0.0023</td>
<td>0.0187</td>
</tr>
</tbody>
</table>

the single-scale and case-wide approaches. While registering IHC slides to H&E stained slides, we found our multi-scale MSER method to perform better as compared to its single-scale counterpart [2]. However, the case-wide approach having used all the slides in a stack was able to find a better sequence along with the stack for performing alignment and hence has performed better than the multi-scale approach. Overall, the multi-scale approach has been shown to improve registration better than the single-scale and case-wide approaches.

The average registration error values obtained for cases 3, 4 and 5 are comparatively higher as compared to that of cases 1, 2, 6 and 7, as reported in Figure 3.7. On analysing these images and estimated transformations, we found that the high registration error was attributed to the H&E vs IHC setting for cases 3 and 4 and other than morphological changes between IHC and H&E images, some part of the tissue was missing in one of the images in a pair. Hence, registration of these cases in the H&E vs IHC setting was challenging. While, for case 5, MMSER was unable to align one pair of images successfully which contributed to the higher error for this case.

To ensure the feasibility of feature matching at magnification level $1.25\times$, we performed feature detection and matching in a divide and conquer manner which can be performed in a parallel fashion. The bottom row in Figure 3.7 shows the processing time for single-scale [2], global transform, tissue transform and block-wise transform. The difference between single-scale and global transform is only the $\Delta$ value; for single-scale, it is 1.5 while for global transform, it is selected automatically for each case. Single-scale transform has a comparatively longer processing time for cases 1 and 7 as compared to that of global transform. This is because a $\Delta$ value higher than 1.5 was selected for global transform, resulting in less number of MSERs. The maximum processing time of single-scale, global, tissue and block-wise transforms is around 34 seconds, 8 seconds, 46 seconds and 3 minutes, respectively.
Figure 3.7: Bar charts demonstrating registration efficacy and processing time for each case individually. Initial refers to the measures computed before alignment while global, tissue and block-wise are the series of transforms proposed for multi-scale MSER based registration (as discussed in section 3.3.3). The top and middle bar charts present registration performance for different steps in our multi-scale approach. Similarly, the comparison of processing time is shown in the bottom row, where single-scale corresponds to [2].
Table 3.3: Comparative results of registration using a single-scale, multi-scale MSER (MMSER) and case-wide approach. This comparative analysis is presented for two different settings: IHC vs IHC and H&E vs IHC. Registering IHC stained slide to H&E slide is challenging due to the differences in tissue architectures.

<table>
<thead>
<tr>
<th></th>
<th>Avg(Med(rTRE))</th>
<th>Med(Med(rTRE))</th>
<th>Avg(Max(rTRE))</th>
<th>Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>IHC vs IHC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSER</td>
<td>0.0031</td>
<td>0.0019</td>
<td>0.0086</td>
<td>0.9994</td>
</tr>
<tr>
<td>MMMSER</td>
<td>0.0021</td>
<td>0.0017</td>
<td>0.0067</td>
<td>0.9991</td>
</tr>
<tr>
<td>Case-wide</td>
<td>0.0053</td>
<td>0.0033</td>
<td>0.0114</td>
<td>0.9801</td>
</tr>
<tr>
<td><strong>H&amp;E vs IHC</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSER</td>
<td>0.0125</td>
<td>0.0111</td>
<td>0.0779</td>
<td>0.9550</td>
</tr>
<tr>
<td>MMMSER</td>
<td>0.0166</td>
<td>0.0072</td>
<td>0.0349</td>
<td>0.9782</td>
</tr>
<tr>
<td>Case-wide</td>
<td>0.0120</td>
<td>0.0073</td>
<td>0.0381</td>
<td>0.9528</td>
</tr>
<tr>
<td><strong>All</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSER</td>
<td>0.0156</td>
<td>0.0028</td>
<td>0.0510</td>
<td>0.9390</td>
</tr>
<tr>
<td>MMMSER</td>
<td>0.0068</td>
<td>0.0023</td>
<td>0.0187</td>
<td>0.9925</td>
</tr>
<tr>
<td>Case-wide</td>
<td>0.0075</td>
<td>0.0049</td>
<td>0.0290</td>
<td>0.9715</td>
</tr>
</tbody>
</table>

3.6 Summary/Discussion

Due to the large size of WSIs, it is attractive to perform registration at a low-resolution version of the WSI. In this case, refining the transformation in a multi-resolution strategy is deemed suitable. In this chapter, we showed that high registration accuracy can be achieved with MSER features when alignment is performed in a multi-scale manner. This is affirmed by the comparative analysis of our multi-scale approach with the single-scale MSER registration approach. Our results demonstrate that the multi-scale approach is capable of aligning the sections better, even in cases where there are significant differences in the tissue structures of an image pair. Although the multi-scale approach improves the registration performance, it comes with a long processing time (max 3 minutes). Performing MSER at a high resolution resulted in a significant rise in the number of MSERs detected. It was likely to have a longer processing time since there were many potential valid pairs. There is a need for a registration method to be time-efficient along with a better performance at the same time.

There are a number of other handcrafted features which have been widely used for matching tasks in various applications such as HOG [93], SIFT [75], SURF [94], ORB [95] and BRISK [96]. However, none of them can perform registration successfully for all the imaging datasets. There has been a lot of interest from the research community to learn features from the data due to the inherent limitations of handcrafted features. Recently, data-driven approaches have been developed to deal with various aspects of alignment such as correspondence detection, transformation estimation etc. We investigate data-driven features for correspondence detection which is the focus of our next chapter.
Chapter 4

Deep Feature Matching based Registration

4.1 Introduction

In the previous chapter, we employed handcrafted features for registering serial sections and demonstrated their inherent limitations for images in the histology domain. We also investigated their efficacy when applied at a series of magnifications, starting from low magnification and moving to immediate high magnification levels. In comparison to the single-scale approach, the multi-scale approach was shown to improve the spatial overlap between image pairs and its success was attributed to extracting a larger number of MSERs at high resolution. However, it has a few limitations that hinder its practicality when applied in a real scenario which requires global transformation to be fast to not interfere with the following non-linear alignment and cross-slide analysis. Images taken from the high magnification of pyramid structure resulted in a higher number of MSERs and took an excessively long processing time. We also observed some cases for which the multi-scale approach failed to give optimal registration. In case of failed transformation estimated by the tissue transform, the following block-wise transform is not likely to improve the alignment further since each pair of tissue blocks would contain different tissue regions. There were two possible reasons for a failed registration: first not getting any matching points that fulfil the conditions and second getting matching points that do not map to the same tissue regions. These limitations can be overcome by employing a feature extractor which generates a fixed number of features irrespective of the image; features that can enable robust feature matching.

Deep learning has emerged to provide an alternative way of learning data learned features instead of manually designing features using the information present in the images [97]. CNNs have been utilised as a feature detector
and descriptor in several studies in recent years and are shown to be more invariant to changes in image appearance [98–100, 100–103], a property vital for multi-modal and multi-stained images. After the ANHIR challenge [1], deep learning has been used mainly for predicting transformation parameters of higher degrees of freedom for registering histology images. However, the option of using deep features in feature matching for direct estimation of linear transformation is limited. To the best of our knowledge, there is only one study performing feature matching using CNN features for the end purpose of registering non-histology images [104]. We adapted this study to estimate rigid transformation parameters for an image pair. Our motivation for employing CNN features is two-fold:

1. To investigate CNN features for registering histology images.
2. To overcome the problem of having a very fewer or an enormous number of features (as observed in the MSER based approach). Using CNN as a feature extractor resulted in a fixed number of features for every image.

The focus of this chapter is on demonstrating the robustness of data-driven features for the multi-IHC tissue registration task. We employed these features for estimation of the rigid transformation parameters and refer to this approach as “deep feature based registration” (DFBR). We present a comparative performance analysis of deep features and hand-crafted features. Experimentation is done on the COMET dataset and an additional multi-IHC histology dataset which was released by the organisers of a recent challenge contest on non-linear registration.

4.2 Literature Review

CNNs have been extensively used for various applications for learning base functions from the data itself. These functions learn both low and high dimensional features from the data and are optimised by comparing the output with the ground truth for any particular task at hand. They have been shown to outperform other classical machine learning approaches in many medical image analysis tasks such as segmentation, classification, regression etc. There has been a significant amount of work on deploying deep learning for the registration of non-histology images as compared to microscopic histology images. In literature, the use of deep learning for the registration has been focused on three different approaches: feature descriptor, feature matching using similarity metrics and predicting the transformation parameters.

CNN features have been shown to outperform handcrafted features by a large margin in several different tasks including registration across the computer
vision and medical imaging domains. Several studies have attempted to use deep learning models as feature descriptors in a matching task for medical images. In [105, 106], the authors proposed an unsupervised approach for learning most discriminating features which are later integrated into the existing registration tools for the prediction of a dense deformation field. This study was tailored to registering MR brain images. In line with this approach, deep features have been used for predicting the rigid transformation for histology images [107]. In that work, an autoencoder was trained to generate an output similar to the input to learn a feature representation. Features extracted from the encoder part of a trained autoencoder were used to find the best transformation using gradient descent.

Our proposed approach for registering multi-stain images comes under the same umbrella of using CNN as a feature descriptor. Our work is inspired by the work in [104] on registering multi-temporal remote sensing images using a CNN, whereby the authors used multi-scale deep features for detection of matching feature points between an image pair. These matching feature points are then used to solve thin-plate spline (TPS) interpolation for image alignment. In our work, we have followed a similar approach for feature description and the correspondence between feature points of two images was found by computing the Euclidean distance measure. Instead of TPS formulation, we used these matching feature keypoints for estimating the rigid transformation.

Previously, there were very few CNN based studies on predicting the transformation parameters for a highly deformed pair of images. This is because the known correspondence is needed for training a CNN which is not often available. Also, the trained CNN may not perform comparably well on an unseen dataset. These limitations can be addressed by using an unsupervised approach to some extent and are yet to achieve a significant improvement in terms of registration accuracy. However, there is a wide gap in the extent of utilisation of the unsupervised deep learning approach for registration tasks as compared to other image analysis tasks. A major reason is the inability to estimate the transformation while training a network. In 2015, Jaderberg et al. [108] proposed a learnable module for applying the spatial transformation to an image, referred to as the ‘spatial transformer’. Since this module is differentiable, it can be added to any network for end-to-end training. This was not primarily designed for registration purposes; instead, the aim was to enable the CNN to learn features that are invariant to the spatial transformations. After the introduction of the spatial transformer, deep learning gained momentum in designing neural network architectures suitable for registration in an unsupervised learning manner. It has now become a core component of most deep learning-based registration methods. Chang et al. [109] proposed a multi-scale iterative framework for registering microscopic
images of serial sections of a mouse brain. A CNN model with a spatial transformer as one of the layers was used to predict the affine transformation. The model was trained to reduce the mean square error between the reference and warped moving images. In [89], the authors proposed a deep learning architecture for 3D registration of CT images of the liver and MR images of the brain. The proposed architecture consisted of two sub-networks: one for predicting the affine transformation and the other for predicting the non-linear deformation field. This approach was also applied to the multi-stain histology dataset provided as a part of the ANHIR challenge [1] and was observed to be the fastest performing method. However, it was not close to the best-performing methods in terms of registration error due to its limited generalisability. It was ranked 6 out of 10 teams who submitted the results. In another study [110], the authors trained a CNN model with good generalisability for predicting the affine transformation in an unsupervised manner. They compared their results with that of SIFT, SURF and Elastix tool. Their proposed approach didn’t outperform other methods in terms of accuracy; however, considering the success rate, the authors claimed that the reported accuracy had been sufficient to perform non-linear alignment successfully. In [111], the authors integrated segmentation maps to aid in performing non-linear registration using a self-supervised deep learning-based approach. Segmentation maps were generated by applying $k$-means clustering to concatenated multi-scale feature maps, extracted from a pre-trained segmentation model. They employed VoxelMorph architecture [112] for registration and replaced the manual segmentation maps with their fine-grained feature maps. Similar to other non-linear registration methods, this method also required two images to be linearly aligned before its application.

Most approaches to non-linear registration are designed to work with images that are linearly aligned beforehand. This is because the non-rigid transformation can be more successful when images are linearly aligned. There has been a significant amount of work on non-linear registration. However, the focus has been on monomodal and multi-modal registration of X-ray, CT and MRI images and very few methods have been proposed for histology images. Wodzinski and Muller [113] proposed a deep learning based non-rigid registration method, performing comparably to the winning team of the ANHIR challenge contest and is significantly faster than other iterative methods. Their proposed approach employs UNET like architecture, trained in a multi-level unsupervised manner using negative normalised cross-correlation (NCC) as an objective function. However, most data-driven approaches require a large number of training samples to perform highly accurate registration.

In this chapter, our main focus is on the development of a pipeline for robust and accurate estimation of the rigid transformation for multi-stain
histology images. Cross-section slides mostly have non-linear deformation that cannot be tackled with rigid transformation. To mitigate these limitations, we employed an existing method for non-linear registration which has been successfully applied to multi-stain histology images.

4.3 The Proposed Approach

A registration step for any downstream colocalisation analysis workflow should be able to allow a significant spatial overlap between the two images such that the location of corresponding tissue structures can be determined. To this end, we propose a pipeline comprising three main steps: pre-processing and estimation of rigid alignment using our proposed DFBR method, followed by a non-linear registration. During the pre-processing step, we generate a tissue mask for an image pair and modify the input images such that they appear spatially similar. Our DFBR method further contains three sub-modules: pre-alignment to perform rough alignment, tissue transform estimates the transformation parameters using cropped tissue region and block-wise transform refines tissue transform by performing feature matching in a block-wise manner. After an image is registered using our DFBR, we observed a slight offset in some cases. To fix this offset, we added a local transform module which is followed by an existing non-linear registration method [84, 85]. The overall proposed pipeline for registration is shown in Figure 4.1.

![Figure 4.1: Overall pipeline for our proposed cross-slide image registration.](image-url)
4.3.1 Pre-processing

Tissue Segmentation

In tissue segmentation (TS₁) used in our basic registration pipeline (section 3.3.3), all tissue and non-tissue regions were considered as foreground and background, respectively. First, we conducted our DFBR experiments using the TS₁ masks. However, we observed that feature matching points from the fatty region were mostly mismatched. This is because the fatty tissue does not contain adequate texture, thus resulting in low discriminating features. To avoid getting incorrect matching points, we modified our ground truth for tissue segmentation. In the updated masks, we considered the fatty region as a background. In this thesis, we will refer to this tissue segmentation as TS₂. In Figure 4.2, a downsampled version of WSI is shown along with its two different tissue masks to demonstrate the difference between the two. We trained a separate CNN for generating these masks so that registration could be carried out using matching points from the active or discriminatory tissue area only (while excluding the ones heavily surrounded by the fatty tissue). Our DFBR method performed better when the transformation matrix was estimated using matching points from the discriminatory tissue area. The quantitative and qualitative comparison of DFBR using TS₁ and TS₂ are shown in Table 4.1 and Figure 4.3. Similar to basic registration, control tissue was excluded while estimating the transformation parameters.

<table>
<thead>
<tr>
<th></th>
<th>AMrTRE</th>
<th>MMrTRE</th>
<th>AMaxrTRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS₁</td>
<td>0.0090</td>
<td>0.0044</td>
<td>0.0208</td>
</tr>
<tr>
<td>TS₂</td>
<td>0.0063</td>
<td>0.0041</td>
<td>0.0176</td>
</tr>
</tbody>
</table>

Table 4.1: Comparative results of our proposed rigid alignment approach (DFBR) using two different tissue segmentation masks. TS₁ refers to an overall tissue region mask including fatty tissue and TS₂ refers to an active tissue region mask excluding fatty tissue. The exclusion of fatty areas significantly reduced the registration error. These results are generated using the COMET dataset.

Image Pre-processing

CNN models trained for a classification problem are known to be biased towards texture rather than the colour of the input image [114]. Based on this observation, we can perform feature matching using original images without any pre-processing. To explore it with our registration pipeline, we experimented with four different versions of the input images: original RGB images, greyscale images, blue ratio and H stain images. To unify the appearance of two images, we performed histogram matching as a normalisation step. In histogram
Figure 4.2: An example image with the two types of tissue masks. In TS_1, all the tissue including fatty region is included whereas in TS_2, fatty region is excluded due to its negative impact on our DFBR registration method.

Matching, the histogram of an image is modified to be similar to that of another image. It was performed for all pairs of images except those with RGB values. In our experiments, an image with high entropy is considered a reference image and the histogram of an image with low entropy is matched to the reference image. Since most pre-trained CNN models accept input images with 3 channels, greyscale, H and blue ratio images were stacked as the colour channels. In our experiments, we found greyscale images to perform better as compared to other input versions, as shown in Table 4.2.

<table>
<thead>
<tr>
<th></th>
<th>AMrTRE</th>
<th>MMrTRE</th>
<th>AMaxrTRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Ratio</td>
<td>0.0145</td>
<td>0.0058</td>
<td>0.0285</td>
</tr>
<tr>
<td>RGB</td>
<td>0.0077</td>
<td>0.0043</td>
<td>0.0190</td>
</tr>
<tr>
<td>H stain</td>
<td>0.0070</td>
<td>0.0044</td>
<td>0.0185</td>
</tr>
<tr>
<td>Greyscale</td>
<td>0.0063</td>
<td>0.0041</td>
<td>0.0176</td>
</tr>
</tbody>
</table>

Table 4.2: Demonstration of the efficacy of our DFBR approach using different versions of input image pairs. Greyscale images are shown to outperform other pre-processed input images. These results are generated using the COMET dataset.
Figure 4.3: A qualitative comparison of the registration accuracy of transformations estimated using segmentation TS₁ and TS₂. The first and second columns in the first two rows show reference and moving images, respectively while the third and last columns show transformed moving images using segmentation TS₁ and TS₂, respectively. i) shows overlaid images of e) and g) while j) shows overlaid images of e) and h). Landmarks are overlaid for the purpose of visualising the offset between the two images.
4.3.2 Alignment

Broadly speaking, the tissue alignment is performed in three main steps: pre-alignment, rigid alignment and non-rigid alignment. The output registered image generated in each step is given as an input to the next step along with the reference image. All these steps are discussed in detail in the following sections.

Pre-alignment

In this step, rough estimates of translational and angular offsets are computed. First, we estimate the translation offset by finding a centre of mass (COM) for an image pair. The COM is a vector of \(x\) and \(y\) coordinates and is computed from the inverted greyscale intensity values of the tissue region only. The difference between the COM values of a pair of images is used to obtain a translation matrix. This matrix will transform the moving image such that its COM is at the same position in the coordinate system as that of a reference image. Next, we find a rotation matrix by an exhaustive search strategy resulting in a maximum DICE overlap between the tissue masks of an image pair. With the rough alignment of the images computed, the moving image is transformed resulting in a reasonable overlap with the reference image. We then crop the tissue regions from a roughly aligned image pair using their tissue masks. We determine a bounding box that includes the tissue region of both images. In the following steps, we employ features matching in the tissue regions of an image pair instead of the whole image for registration. The matched features provide a set of spatial correspondences for better alignment.

Deep Feature based Registration

The objective of our feature based alignment step is to refine the alignment between reference and pre-registered moving images by registering their feature points. In Chapter 3, we used MSER features for this purpose. Here, we present data-driven features extracted using a pre-trained VGG-16 model. The partial architecture of VGG-16 that we used for feature extraction is shown in Figure 4.4. We extract multi-scale features for an image pair and find the matching pairs by considering the feature points with a small feature distance.

A. Feature Descriptors: In a handcrafted feature based method for registration (as in the previous chapter), the first step is to identify/detect the feature points containing distinctive information such as a corner, blobs, edges etc. It is then followed by a feature extraction step which involves the computation of descriptors. Descriptors describe the properties of regions centred at the distinctive feature points. These feature descriptors are used to find correspondence between two images in a feature matching step. In
Figure 4.4: A partial architecture of the VGG16 model, used for feature extraction in our DFBR module.

Our proposed feature-based method for alignment, the feature detection step doesn’t exist. Instead, an image is divided into a grid and a feature descriptor is computed for every grid cell in a sliding window fashion. Similar to [104], our feature descriptor is formed by deep features extracted from three different layers of a VGG-16 model, pre-trained for ImageNet classification. Since we have removed the bottom fully connected layers, the partial model can take input images of any size (spatial dimensions multiples of 32), with larger images increasing the computation time. Our experiments are conducted with an input size of 224 × 224 pixels. Features extracted from the bottom three pooling layers (pool3, pool4 and pool5) are used to build descriptors. Each of these layers corresponds to receptive fields of different sizes as shown in Figure 4.5. Pool3 is considered a reference feature extractor layer; therefore, the features of pool4 and pool5 are mapped to those of pool3. Following are some annotations that we use in this section: \( F^j_i \) refers to the features of an image \( j \) extracted from a pooling layer \( i \). For example, pool3 layer features of a reference image \( R \) is denoted by \( F^R_3 \).

Figure 4.5: Demonstration of feature summation for a sub-image.
• **Pool3** outputs features $F_3$ of dimension $28 \times 28 \times 256$. Each feature has a receptive field of size $8 \times 8$, dividing input image into a $28 \times 28$ grid. The centre of each grid cell is considered a feature points for the respective descriptor.

• **Pool4** outputs features $F_4$ of dimension $14 \times 14 \times 512$. Each feature has a receptive field of size $16 \times 16$ and therefore it is shared by four feature points.

• **Pool5** outputs features $F_5$ of dimension $7 \times 7 \times 512$, each feature has a receptive field of size $32 \times 32$ and therefore it is shared by 16 feature points.

Each of these descriptors is normalised to unit variance.

**B. Feature Matching:** Once feature descriptors are formed for an image pair, we compute the Euclidean distance between their feature points. The distance metric for a layer $i$ features is computed as

$$D_i(p^R, p^M) = \text{distance}_{\text{euc}}(F_i^R, F_i^M)$$

where $i \in [3, 4, 5]$. Each value in a feature distance matrix for pool3 relates to an individual feature point, which is not the case with distance matrices for pool4 and pool5 features. Each distance value in $D_4$ and $D_5$ corresponds to 4 and 16 feature points, respectively. Therefore, we replicate each distance value in $D_4$ 4 times. Similarly each distance value in $D_5$ is replicated 16 times. After replication, $D_3$, $D_4$ and $D_5$ are added, with some weight given to $D_3$ due to a smaller number of features. Feature distance between two feature points $p^R$ and $p^M$ is computed as

$$D(p^R, p^M) = \sqrt{2}D_3(p^R, p^M) + \text{replicate}(D_4(p^R, p^M), 4) + \text{replicate}(D_5(p^R, p^M), 16)$$

The quality of each matching feature point is determined by the difference between the smallest and second smallest Euclidean distances. Greater the difference, the better the quality. An example illustration for this is shown in Figure 4.6.

There are two conditions for matching point $p^R$ to point $p^M$

1. $D(p^R, p^M) < D(:, p^M)$; which means there shouldn’t be any other feature distance smaller than $D(p^R, p^M)$.

2. The quality of matched features should be greater than a threshold value which is computed automatically for each image pair. A threshold value is set such that $U$ number of matching points are selected.
Figure 4.6: Demonstration of how quality of matched features is determined. An example distance matrix computed between the CNN features of two images is shown. Features with the smallest distance $min_1$ are considered as matched features. $min_2$ is the second smallest distance after $min_1$. The quality of matched features is determined by the difference between $min_1$ and $min_2$.

Using the above steps, we find matching feature points between any two images. Similar to [104], we select $U = 128$ pairs of matching points. We then use the matching feature points as control points for estimating the transformation parameters. Using a least squared approach for finding a rigid transformation that best aligns these matching points. To use this method for histology image registration, we apply this step twice. First, we apply it to the whole tissue region for finding the best matching points globally which is referred to as the ‘tissue transform’. Secondly, we divide the tissue regions into four parts and apply our feature matching method to each part individually. This is referred to as the ‘block-wise transform’. It results in a higher number of matching feature points and is likely to further improve the alignment. Also, the block-wise feature matching step can be applied in a parallel fashion to speed up the process.

Local Alignment

On visualising a registered image alongside its corresponding reference image at both low and high resolution, we observed global and local translation offsets, respectively. The global offset was likely to exist due to the offset in patches containing the matching keypoints (centre of the patch). On viewing an image pair at a higher resolution with a minor global offset, we observed a local translation offset between reference and registered images. This is likely to persist due to the non-linear deformations in some parts of the images. Figure 4.7 shows an overlaid image of reference and registered moving tissue sections along with their corresponding landmark points. The non-overlapping landmarks demonstrate the local offset between the two images. With landmark
points, it can be observed that some parts of the image pair are properly aligned where landmarks of the two images are overlapping, while regions with distant corresponding landmarks indicate a local translative movement between the two images. To fix a global offset, we employed the phase correlation method [115] after applying DFBR to determine a shift between two images at magnification $0.3125 \times$. It is a fast alignment method suitable when one image is translated relative to another image and does not present any other transformations [116]. We found this method performs better when applied to tissue masks, however it can be applied to the greyscale or H stain images. We also integrated this method in our visualisation tool (section 4.4) to fix the local shift. During visualisation, the user can fix the offset by clicking on the ‘Fix Offset’ button. An example of an image pair before and after local refinement is shown in Figure 4.7.

**Non-Rigid Alignment**

Once an image pair is registered using rigid transformation, it is often the case that some of the tissue areas are not accurately aligned to that of a reference image. As an example, an overlaid image of reference and registered images is shown in Figure 4.8. The landmarks are also displayed over images. It can be observed that most of the landmarks are accurately overlapping while there are some landmarks that are not registered well. This is due to the fact that the tissue slices are so thin and fragile that the slide preparation step is likely to introduce non-linear deformations and artefacts such as tissue folds, tissue stretching and compression and even torn/missing tissue parts [117]. The presence of non-linear deformations makes the registration process more challenging. Since these artefacts change the morphology of the tissue, none of the rigid registration methods can tackle such deformities. Therefore, a non-rigid registration approach is applicable in such scenarios. It should be noted that the non-rigid transformation can be applied successfully when two images are linearly aligned. Otherwise, it can introduce further artefacts in the transformed image.

There are many existing non-rigid registration methods in the literature which have been used for histology image alignment. In a recent challenge on the alignment of multi-stain histology images, a number of non-rigid methods were evaluated using a benchmark dataset. We have collaborated with the MEVIS group (winner of the ANHIR challenge) for improving our registration results further by applying their non-rigid algorithm [84, 85] on top of our DFBR method. Their proposed method as used in the challenge comprises three steps: pre-alignment, parametric registration and non-parametric registration. Their pre-alignment step is similar to our pre-alignment step except that their
Figure 4.7: Example images of the local phase-based refinement method integrated into our visualisation tool (see section 4.4) to fix translation offset. In bottom row g) we show overlaid false colour images before (left) and after refinement (right). Reference and registered moving images are shown in green and purple colours, respectively.
method does not incorporate the tissue mask while computing the centre of mass. The NGF distance measure [86] is used for the selection of the best alignment in this step and the following steps. The NGF measure is computed using the following formulation:

\[ NGF(R, M, y) := \frac{h^2}{2} \sum_{i=1}^{N} 1 - \left( \frac{\langle \nabla M(y(x_i)), \nabla R(x_i) \rangle_\varepsilon}{\| \nabla M(y(x_i)) \|_\varepsilon \| \nabla R(x_i) \|_\varepsilon} \right)^2 \]

(4.1)

with \( \langle x, y \rangle_\varepsilon = x^\top y + \varepsilon^2 \) and \( \| x \|_\varepsilon := \sqrt{\langle x, x \rangle_\varepsilon} \). \( x_i \) represents pixel centers (i = 1, ..., N), \( y \) is the transformation and \( \varepsilon \) is responsible for controlling the sensitivity of edges in relation to noise. The NGF value reduces when two images are aligned.

In the parametric registration step, an affine transformation is estimated based on the greyscale intensity values using the NGF distance measure. A Gauss-Newton optimisation is used to minimise the objective function

\[ J(R, M, y) = NGF(R, M, y) \rightarrow \text{min} \]

(4.2)

where \( y \) is the estimated affine transformation. These transform parameters are then set as an initialiser for the non-parametric dense registration which is
given by

\[ y(x) = x + u(x) \quad (4.3) \]

with displacement \( u : \mathbb{R}^2 \to \mathbb{R}^2, u = (u_1, u_2) \) [118]. The displacement field is estimated by minimising their proposed objective function comprising the NGF distance measure along with a curvature regulariser which is defined as

\[
CURV(y) = \frac{h^2}{2} \sum_{i=1}^{N} \left| \Delta^h u_1(x_i) \right|^2 + \left| \Delta^h u_2(x_i) \right|^2
\quad (4.4)
\]

with 2D Laplacian \( \Delta = \partial_{xx} + \partial_{yy} \), \( \Delta^h \) is a five point finite difference approximation of \( \Delta \) with Neumann boundary conditions and \( y \) is the non-linear deformation. L-BFGS is used for optimisation and the objective function is minimised using

\[
J(R, M, y) = NGF(R, M, y) + \alpha CURV(y) \rightarrow \min \quad (4.5)
\]

where \( \alpha \) is a regularizer to control the smoothness of the deformation. The efficacy of computed deformation is highly dependent on the accuracy of the initialiser transform. The size of the control point grid is significantly smaller than the resolution of the image and is of size \( 257 \times 257 \). For more details on this non-linear registration approach and parametric values, readers are referred to [84].

We have experimented with the method proposed by the MEVIS group and found some weaknesses in their pre-alignment step. First, transformation is estimated based on the whole slide content rather than the tissue region only. Due to the existence of slide artefacts and control tissue, we consider tissue segmentation as an integral pre-processing step. Secondly, the parametric alignment step uses intensity values rather than key features for transformation estimation. Due to the aforementioned reasons, we replaced the first two steps of the MEVIS method with our DFBR method.

### 4.4 Visualisation Tool

To visualise registered images while being able to zoom in and out and pan across a WSI, we developed two different web-based tools with a split screen, the left panel for displaying the reference image and the right panel for the registered moving image. On each split screen, a dot pointer with a different colour is shown which changes its position with the mouse movement. The regions pointed by these points on two screens are significantly helpful in visually estimating the performance of the registration method. A screenshot of the interface is shown in Figure 4.9.
Figure 4.9: Visualisation Tool. Reference and registered moving images are displayed on the left and right side panels, respectively. Zooming and panning on these panels are interlinked. It allows the user to visualise corresponding regions of two registered WSIs at any resolution level.
**Interface-1** is developed to display the results of our deep feature based method. The input to this tool comprises three directory paths: to reference and moving images and the pre-computed transformation parameters. Registration is applied to the tiles on the fly as they are viewed. Therefore, there is no need to generate transformed WSI in a pyramidal format. Since registration was performed at the lowest resolution, we observe a translation offset between image pairs. It may also be due to the non-linear transformation. In this case, a partial tissue is distorted such that no global rigid transformation can fix the whole tissue. We dealt with such misalignment by using the phase correlation method. We added a button to the interface for the user to fix the offset. Once the offset is computed, it is applied to every field of view (FOV) as the user zoom or pans through the slide.

**Interface-2** is developed to display results of non-rigid registration method. It is purely a visualisation tool with no operation on tiles at the back-end while the user is visualising them. This is developed to visualise reference and registered images saved in a pyramidal format. After non-rigid registration, we save the registered images as multi-scale images and use this interface to assess the performance visually. This interface takes two inputs only, the path to reference and transformed moving images.

### 4.5 Datasets & Performance Measures

We evaluated the performance of our proposed DFBR method using two multi-stain datasets: the COMET and ANHIR datasets. A detailed description of the datasets and evaluation metrics is presented in the following subsections.

#### 4.5.1 Datasets

Details of the COMET dataset are provided in the previous chapter (section 3.4.1). The ANHIR dataset is a public dataset, made available by organisers of the ANHIR challenge. It comprises 8 different tissue types stained with 18 different stains, hence making it a challenging dataset. Example images are shown in Figure 4.10. There are 230 training and 251 testing pairs for registration. Summary of this dataset is presented in Table 4.3. For more details on this dataset, readers are referred to the challenge paper [1]. The landmarks are provided for the training set while landmarks of reference images in test image pairs are kept private by the organisers. The evaluation of a registration method can be performed by uploading the results on the challenge portal.
Figure 4.10: Example images from the ANHIR dataset. Each image is stained with a different stain.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Scanner</th>
<th>Mag.</th>
<th>microns per pixel</th>
<th>Avg. dimensions</th>
<th>Number of train pairs</th>
<th>Number of test pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lung lesions</td>
<td>Zeiss</td>
<td>40×</td>
<td>0.174</td>
<td>18k × 15k</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>Gastric</td>
<td>Leica</td>
<td>40×</td>
<td>0.2528</td>
<td>60k × 75k</td>
<td>13</td>
<td>40</td>
</tr>
<tr>
<td>Human breast</td>
<td>Leica</td>
<td>40×</td>
<td>0.253</td>
<td>65k × 60k</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Human kidney</td>
<td>Leica</td>
<td>40×</td>
<td>0.253</td>
<td>18k × 55k</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Mouse kidney</td>
<td>Hamamatsu</td>
<td>20×</td>
<td>0.227</td>
<td>37k × 30k</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>Lung lobes</td>
<td>Zeiss</td>
<td>10×</td>
<td>1.274</td>
<td>11k × 6k</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Mammary glands</td>
<td>Zeiss</td>
<td>10×</td>
<td>2.294</td>
<td>12k × 4k</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>COAD</td>
<td>3DHistech</td>
<td>10×</td>
<td>0.468</td>
<td>60k × 50k</td>
<td>84</td>
<td>153</td>
</tr>
</tbody>
</table>

Table 4.3: Details of the ANHIR dataset [1].

4.5.2 Performance Measures

To measure the quality of our registration pipeline, we compute the TRE using landmarks of reference R and registered moving M images, as described in Section 3.4.2.

4.6 Experimental Results

4.6.1 Experimental Setup

We considered different scales of information in our registration pipeline. For the COMET dataset, downsampled WSIs at magnification 0.15625×, 0.3125× and 0.625× were used for pre-alignment, DFBR including local transform and
non-rigid alignment, respectively. For the ANHIR dataset, downsampled WSIs were provided rather than WSIs with varying downsampling rates for different datasets. We rescaled the images to 5% and performed our DFBR registration, followed by a non-rigid registration. For non-rigid registration, we used the same parametric values as the challenge winner. These parametric values are reported in Table 4.4. The python based library Keras was used for the extraction of CNN features while the rest of the implementation is in both Python and MATLAB. The implementation of our visualisation tool is carried out in Python and JavaScript. We used OpenSeadragon, an open-source viewer for this tool.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of levels</td>
<td>7*</td>
</tr>
<tr>
<td>Maximum image dimension</td>
<td>8000 pixels</td>
</tr>
<tr>
<td>NGF $\varepsilon$</td>
<td>1.0</td>
</tr>
<tr>
<td>Regulariser parameter $\alpha$</td>
<td>0.1</td>
</tr>
<tr>
<td>Control point grid size</td>
<td>$257 \times 257$</td>
</tr>
</tbody>
</table>

Table 4.4: Parameters of non-rigid registration method: $\alpha$ is a regularisation parameter to control the smoothness in deformation while $\varepsilon$ controls the sensitivity of NGF to the noise, and * indicates the parameter that differs in the COMET dataset. We performed registration at a total of 6 levels for the COMET dataset.

### 4.6.2 Result Summary

**COMET:** As discussed previously, we observed mismatched feature points using our DFBR method, mainly in the fatty tissue area. Therefore, we replaced our tissue mask TS$_1$ with TS$_2$. The only difference between the two is we have excluded the fatty region in TS$_2$. The removal of mismatched feature points from the fatty area improves our results, as shown in Table 4.1. Table 4.2 shows results obtained with different versions of input images: original RGB images, greyscale images and H stain images. As our initial experiments showed greyscale images to perform better, greyscale images are used in all our experiments.

Each step in our pipeline was adopted to improve the alignment. The quantitative results in Table 4.5 demonstrate that our first three steps result in improvement of the rigid alignment between the two images while the last step has been able to tackle the non-linear deformation. Box plots shown in Figure 4.12 demonstrate the reduction of median rTRE after each step in our pipeline. The pre-alignment step has significantly improved the median rTRE. Figure 4.13 shows overlaid false colour images of reference and registered moving images along with the zoomed in visual fields. In overlaid images, the extent
of the green colour indicates the extent of misalignment. It can be seen that the misalignment is improved after each step.

In Figure 4.11, average registration errors are shown for each pair on our evaluation data. CK8/18, MLH1, MSH2, MSH6 and PMS2 appear one after the another in the same sequence. Whereas, H&E appears after PMS2, with 6 sections in between the two. The bar plots in Figure 4.11 show that the registration error typically increases with the distance between the pair of images.

![Bar plots for different reference images](image1)

![Bar plots for different reference images](image2)

Figure 4.11: Quantitative results of our proposed framework on the COMET dataset. These results are presented for all fifteen pairs of images used for evaluation. A bar plot is presented for each reference image (H&E, CK8/18, MLH1, MSH2) separately, except MSH6. The registration is performed to align PMS2 w.r.t. MSH6 with 0.0025, 0.0026 and 0.0073 as $\text{Avg}(\text{Med}(rTRE))$, $\text{Med}(\text{Med}(rTRE))$ and $\text{Avg}(\text{Max}(rTRE))$, respectively.

<table>
<thead>
<tr>
<th>AMrTRE</th>
<th>MMrTRE</th>
<th>AMaxrTRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td>0.1797</td>
<td>0.1068</td>
</tr>
<tr>
<td>Pre-alignment</td>
<td>0.0109</td>
<td>0.0087</td>
</tr>
<tr>
<td>Tissue Transform</td>
<td>0.0073</td>
<td>0.0046</td>
</tr>
<tr>
<td>Block-wise Transform</td>
<td>0.0063</td>
<td>0.0041</td>
</tr>
<tr>
<td>Non-rigid Transform</td>
<td>0.0031</td>
<td>0.0014</td>
</tr>
</tbody>
</table>

Table 4.5: Quantitative results generated using the COMET dataset. These results demonstrate that the average error is reduced with each alignment step in our pipeline.
Figure 4.12: Demonstration of the median rTRE before and after applying each registration module using a box plot. The median rTRE is computed for the COMET dataset.

**ANHIR Dataset:** Similar to the COMET dataset, we evaluated each step of our registration pipeline on the ANHIR dataset. As the organisers made landmarks of the training set publicly available, we first evaluated our pipeline on the training set only. The improvement in median rTRE (training set only) after each step of our method is shown in Figure 4.14. To evaluate the performance of our method on the whole dataset, we made a submission to the challenge portal. In Table 4.6, we show values of different metrics computed for the MEVIS and our proposed methods. Our results compare favourably in terms of aggregation average, median rTRE, maximum rTRE and robustness for the training set. We compare our results to those of the MEVIS team to show the effectiveness of the proposed method.

<table>
<thead>
<tr>
<th>Method</th>
<th>Average rTRE</th>
<th>Median rTRE</th>
<th>Max rTRE</th>
<th>Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Median</td>
<td>Average</td>
<td>Median</td>
</tr>
<tr>
<td>MEVIS_Train</td>
<td>0.0061</td>
<td>0.0030</td>
<td>0.0049</td>
<td>0.0019</td>
</tr>
<tr>
<td>Ours_Train</td>
<td>0.0060</td>
<td>0.0029</td>
<td>0.0049</td>
<td>0.0019</td>
</tr>
<tr>
<td>MEVIS_Eval</td>
<td>0.0044</td>
<td>0.0025</td>
<td>0.0049</td>
<td>0.0018</td>
</tr>
<tr>
<td>Ours_Eval</td>
<td>0.0046</td>
<td>0.0026</td>
<td>0.0041</td>
<td>0.0017</td>
</tr>
<tr>
<td>MEVIS_All</td>
<td>0.0052</td>
<td>0.0030</td>
<td>0.0059</td>
<td>0.0018</td>
</tr>
<tr>
<td>Ours_All</td>
<td>0.0055</td>
<td>0.0029</td>
<td>0.0040</td>
<td>0.0018</td>
</tr>
</tbody>
</table>

Table 4.6: Quantitative results of the winning team and our pipeline on the ANHIR dataset. First row in the header represents the aggregation method for an image pair and second row represents the aggregation method for all pairs in a set.

### 4.6.3 Comparative Results

In Table 4.7, we compare our DFBR method with single-scale MSER, multi-scale MSER (MMSER) and single-scale case-wide approaches on the COMET dataset as presented in Chapter 3. The results are presented for three different
Figure 4.13: Demonstration of the efficacy of our proposed rigid registration in a step-wise manner using overlaid images of reference and registered moving images from the COMET dataset. The top row shows reference and pre-aligned moving images. The middle row shows the overlaid images of the corresponding tissue region: left, middle and right overlaid images are shown as the output of the pre-alignment, the tissue based alignment and the block-wise alignment, respectively. The bottom row shows the zoomed in overlaid patches. Reference and registered moving images are shown in green and purple colours, respectively.
Figure 4.14: Demonstration of the median rTRE before and after each registration module using a box plot. The median rTRE is computed for the ANHIR training set only.

Typically, tissue sections of around 3-5 microns in thickness are sliced from the tissue block. We do not have exact figures related to the thickness of the COMET sections. However, if we consider it to be 5 microns then the spatial distance between the H&E and IHC slides would range between 30-50 microns whereas for IHC images it would be 5 to 15 microns, making the registration of the H&E and IHC slides challenging not just because of staining differences but also due to the morphological differences between them. Our DFBR approach outperforms both single-scale MSER and case-wide approaches for both sets. However, MMSER and DFBR are shown to outperform for the first (IHC vs IHC) and second (H&E vs IHC) sets, respectively. The results shown for the second set demonstrate that our DFBR approach can align images even when the tissue structures vary significantly between two images.

<table>
<thead>
<tr>
<th></th>
<th>Avg(\text{Med(rTRE)})</th>
<th>Med(\text{Med(rTRE)})</th>
<th>Avg(\text{Max(rTRE)})</th>
<th>Robustness</th>
</tr>
</thead>
<tbody>
<tr>
<td>IHC vs IHC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSER</td>
<td>0.0031</td>
<td>0.0019</td>
<td>0.0086</td>
<td>0.9904</td>
</tr>
<tr>
<td>MMSER</td>
<td>0.0024</td>
<td>0.0017</td>
<td>0.0007</td>
<td>0.9994</td>
</tr>
<tr>
<td>Case-wide</td>
<td>0.0053</td>
<td>0.0035</td>
<td>0.0114</td>
<td>0.9801</td>
</tr>
<tr>
<td>DFBR</td>
<td>0.0029</td>
<td>0.0028</td>
<td>0.0091</td>
<td>0.9948</td>
</tr>
</tbody>
</table>

|          |                        |                        |                        |            |
| H&E vs IHC|                        |                        |                        |            |
| MSER     | 0.0425                 | 0.0111                 | 0.0749                 | 0.9589     |
| MMSER    | 0.0166                 | 0.0072                 | 0.0419                 | 0.9782     |
| Case-wide| 0.0129                 | 0.0073                 | 0.0318                 | 0.9529     |
| DFBR     | 0.0114                 | 0.0102                 | 0.0344                 | 1          |

|          |                        |                        |                        |            |
| All      |                        |                        |                        |            |
| MSER     | 0.0159                 | 0.0028                 | 0.0310                 | 0.9790     |
| MMSER    | 0.0068                 | 0.0023                 | 0.0187                 | 0.9825     |
| Case-wide| 0.0093                 | 0.0040                 | 0.0209                 | 0.9115     |
| DFBR     | 0.0063                 | 0.0041                 | 0.0176                 | 0.9965     |

Table 4.7: Comparative results of registration using a single-scale, multi-scale MSER (MMSER) and case-wide approach on the COMET dataset. This comparative analysis is presented for two different settings: IHC vs IHC and H&E vs IHC. Registering IHC stained slide to H&E slide is challenging due to the differences in tissue architectures.
4.7 Summary and Discussion

In the previous chapter, we presented a hand-crafted feature matching method for the registration of multi-IHC images. In this chapter, we presented a deep feature matching approach which was shown to outperform the hand-crafted feature based approach. In the DFBR framework, we introduced a pre-alignment step which produces a roughly aligned image pair. This step would make no difference for MSER based registration due to their rotation invariant property. Although CNN features are not rotation invariant, this step is key to performing deep feature matching successfully. Like the MSER based registration method, tissue segmentation is an essential pre-processing step to exclude feature points from the texture sparse region. Our experiments with deep features showed that a good alignment is difficult to obtain in situations when the slide has more fatty tissue. Tissue masks used by MSER in the previous chapter considered tissue region including the fatty region as a foreground but were not considered suitable here. Therefore, a good tissue segmentation which considers the fatty region as a background is required. It should be noted that this is only related to the DFBR approach and not the MSER based registration.

In a digital stack of tissue sections, we will likely observe differences in the tissue structures across the whole stack. This is due to the thickness of each section in relation to the size of the tissue structures. Smaller structures like cells are likely to be seen between the consecutive sections. However, as the distance between the two sections increases, the same cells and even some bigger tissue structures (such as glands) will not persist. A registration method should be able to perform alignment in the presence of architectural differences between the two sections. Our deep feature method has been shown to outperform all other methods when the distance between image pairs is around 50 microns.

In terms of processing time, the DFBR approach is significantly faster as compared to the MMSER feature based approach while comparable to the single-scale MSER for registering a pair of images. For any image pair, the DFBR approach takes a consistent amount of time with a mean processing time of 14 seconds within 1 standard deviation across 105 pairs of images. While with MSER, registration took 17 ± 13 seconds and with MMSER, registration time varied between 0.5 and 3 minutes. Similarly, processing time varies for case-wide registration; on average it takes around 5 minutes for transforming the whole stack of images [2]. The time consistency with DFBR is because there are a fixed number of feature points using which feature matching is performed. Whereas, the processing time for the MSER based approaches is dependent on the number of MSERs detected. In terms of both the quality of
registration and processing time, we chose DFBR for performing registration of different pairs of multi-stain images which are analysed for MMR status followed by the prediction of MSI in the next chapter.
Chapter 5

Assessment of the DNA Mismatch Repair Status from Histology Images

After registering a stack of multi-stain tissue slides, one can perform any number of cross-slide analyses such as stain co-localisation. In reference to IHC WSIs, stain co-localisation refers to analysing stain expression of multiple protein biomarkers in the same tissue region. In this chapter, we present our work on predicting MMR protein expression and finally the MSI status. We predict MSI status using only a single slide either stained with H&E or cytokeratin 8 (CK8) paired with cytokeratin 18 (CK18), referred to as CK8/18, we call them target slides in this chapter. During the training of our CNN model, we adopted stain co-localisation for incorporating MMR expression locally. To this end, we utilised the IHC slides of MMR markers registered to the target slide. We first provide an introduction to the clinical background of the problem, followed by existing literature on MSI prediction. We then discuss our localised approach for generating training data and its comparative analysis to the unlocalised (weakly supervised) approach in which WSI label is assigned to all its patches. We also investigate the generalisability of our trained model on an external dataset.

5.1 Introduction

Our immune system acts as a natural guard against disease and infection-causing factors. It is very important particularly for cancer patients and is investigated for several reasons: its ability to defend the body against the cancer cells, cancer can affect the normal functionality of the immune system or in worst cases, immune system functionality deteriorates as a side effect of the cancer treatment. The principal component of the immune system is the white
blood cells, also known as immune cells, made in our bone marrow. These immune cells can identify cancerous cells and kill them to prevent their growth. Tumour-infiltrating lymphocytes (TILs) are one of the kinds of immune cells that have been shown to exist in close proximity to cancer cells. Their presence indicates that the immune system is functional and responding against the tumour [119]. However, cancer cells can spread to the bone marrow and weaken the immune system by affecting the production of these blood cells. To boost our immune system in fighting against cancer cells, there is a type of biological treatment known as Immunotherapy. It may help in the long-lasting revocation of cancer by training the immune cells to remember disease cells and fight if they ever come back. The studies have shown that the immune system may retain the benefits of immunotherapy even after the treatment has ended [120].

Immunotherapy has been proven to be a compelling treatment for some cancers when other traditional treatments such as chemotherapy become resistive. However, the number of patients who would benefit from this treatment is limited. There are some predictive biomarkers for the selection of patients who are sensitive to this treatment. MMR status of the cancer is one of the most favourable biomarkers. MMR deficiency in the tumour is correlated with the efficacy of immunotherapy due to a high level of mutations [8]. This correlation has been extensively studied in CRC and similar findings are observed in other tumour types as well. In this chapter, we present our work on identifying MMR expression for the end goal of predicting immunotherapy efficacy.

The MMR system is a highly sophisticated system for maintaining the integrity of DNA during replication and recombination and hence responsible for genomic stability [121]. It comprises MMR proteins which are dependent on four genes: mutL homologue 1 (MLH1), mutS homologue 2 (MSH2), mutS homologue 6 (MSH6) and postmeiotic segregation increased 2 (PMS2). These proteins repair the DNA by identifying errors occurring at microsatellites which are short repetitive DNA sequences. The differences in the length of the microsatellite between normal and tumour tissues of the same patient would indicate the deficiency in MMR and would lead to mutations of the above-mentioned genes. Tumours with variation in microsatellite length are known as MSI tumours and the ones with a high level of variations are termed MSI high (MSI-H). There are two main methods for assessing the MMR status of the tumour: 1) IHC testing which is mainly used in clinical settings to detect MMR proteins and 2) PCR based methods for testing MSI. For PCR testing, National Cancer Institute (NCI) workshop in 1997 recommended following the Bethesda guidelines [122] on the five microsatellite markers known as the Bethesda markers. These markers consist of three dinucleotide loci (D17S250, D2S123 and D5S346) and two mononucleotide loci (Big Adenine Tract [BAT]-25 and BAT-26). Tumours showing instability at two or more markers are
considered as MSI-H and when only one of the Bethesda markers is unstable then the tumour is interpreted as MSI-L. If none of the markers is unstable then the tumour is considered microsatellite stable (MSS). Both IHC and PCR testing methods have false-negative rates and are hence complementary to each other.

In our work, MMR expression status is determined by IHC for MMR proteins. The criterion for IHC testing is straightforward which is the expression or complete loss of staining rather than scoring the intensity of staining as in ER/PR or HER2 IHC testing. In terms of loss expression, MLH1 and MSH2 form a pair with PMS2 and MSH6, respectively. Loss of MLH1 or MSH2 protein makes its partner unstable and hence would be accompanied by loss of expression of their respective partners. However, loss of expression of MSH6 or PMS2 will not affect their partners since they can bind to other molecules to maintain their stability [123]. In CRC, the DAB signal is observed in the nuclei of tumour cells only, for all the four MMR IHC assays. Nuclear staining in normal cells (lymphocytes, fibroblasts or epithelium) is considered the internal positive control. A trained pathologist would assign a clinical status (Intact or Loss) to each case after evaluating the presence or loss of MMR staining. **Intact** status is assigned when an IHC assay shows the presence of DAB stain such that nuclear staining of tumour cells should at least have a similar or stronger intensity to that of internal positive control. The status of **Loss** protein expression is assigned when loss of DAB stain is detected in the tumour nuclei such that DAB staining intensity in cancerous nuclei is absent or less than that of the internal positive control. Identifying all four proteins as Intact indicates proficient MMR (pMMR) status while loss of expression in any one of the proteins indicates deficient MMR (dMMR) status.

In this chapter, our focus is on the detection of CRC cases with dMMR/MSI status which is likely to respond well to immunotherapy as compared to patients with pMMR/MSS status [124, 125]. Recently, there has been some deep learning-based work on MSI prediction using H&E images, as discussed in the next Literature Review section. In these studies, CNNs are trained using MSI/MSS labels while in our work we have utilised MMR labels for training purposes. MSI cases can have different patterns of MMR expression, hence, MMR labels can be considered as fine-grained labels of the coarse-grain MSS/MSI labelling. By providing fine-grain MMR labelling, we show that our approach improves the performance of MSI/MSS classification in comparison to previous approaches.

We identified cases with different patterns of MMR expression using two different target slides independently. These images were accompanied by their MMR expression status as the ground truth labels assigned by the pathologist, along with the corresponding MMR IHC WSIs which were used for defining
their expression status. The ground truth labels were provided at the WSI level and we had no information on which patches are associated with the label. This problem could be studied using weakly supervised learning approaches such as multiple-instance learning (MIL) to find out relevant patches from the WSIs contributing to predicting the final multi-output (MMR status). Using the MIL paradigm, the WSI label is assigned to all its patches and the training can be performed in either an end-to-end manner [126, 127] or in two stages or models [128, 129]. An end-to-end training typically requires a sufficiently large dataset, as in [126] around 44K WSIs are used for training the model and in [127], 10K slides are used, just to list a few. However, in a more recent study, a novel weakly supervised pipeline is proposed for the prediction of MSI status and is shown to perform well using a small dataset comprising of less than 300 slides [130]. In our study, we have 6 times fewer slides for training as compared to [130]. Due to the small amount of dataset, we opt for a two-stage pipeline: first predicting the patch-level scores followed by aggregating these outcomes for predicting WSI labels. Using the MIL approach, we cannot utilise MMR IHC slides which can present useful information in addition to the target image. To make use of MMR slides, we generate patch-level labels without pathologists’ input using registered target and MMR slides, rather than assigning WSI MMR labels to all the patches of corresponding target slides as in MIL methods. This can aid in learning MMR and MSI related biology from target images for interpretation and new knowledge discovery. The comparison of our proposed approach and the two-stage MIL is presented in the Results section.

5.2 Literature Review

MSI testing incurs additional cost and time and hence it is not performed for every patient in a clinical routine. Recently, researchers have been able to predict the MSI status directly from the H&E stained slides using deep learning and their encouraging results demonstrate that such an automatic system could enable MSI screening for every patient worldwide. Kather et al. [131] employed a trained CNN to classify tumour patches in a WSI into MSI or MSS class and finally, the class with a higher number of patches was used for assigning the WSI label. They later refined and tested this system on a large-scale data [132, 133] and in [134], they perform MSI detection using the same approach without performing tumour segmentation as a preprocessing step.

Schmauch et al. [127] proposed a deep learning-based pipeline for predicting RNA-Seq profiles from H&E stained WSIs. The proposed pipeline comprises two main steps. First, they perform CNN based feature extraction for randomly
selected 8000 tissue tiles, resulting in a feature map of size 8000×2048 for each WSI. In the second step, they train a multi-layer perceptron referred to as HE2RNA for predicting gene expression for each tile. During the training stage, they take an average of different numbers (k) of the highest tile predictions at every iteration for deciding slide-level labels. At the inference stage, the slide-level label is obtained by taking an average of all tile-level labels obtained with every possible value of k. For MSI prediction, they simplified the above approach by averaging CNN features for every tile resulting in a feature map of size 1×2048 for every WSI and also modified the HE2RNA model accordingly.

Bilal et al. [130] proposed a novel deep learning framework for predicting the status of several molecular pathways, including MSI. Their proposed framework comprises three models: the first model (ResNet18) performs tumour segmentation, the second model referred to as iterative draw-and-rank sampling (IDaRS) selects the tiles with the most discriminatory features and the last model (HoVer-Net [135]) performs nuclei segmentation and classification into five different cell types. To predict slide-level labels, a linear support vector machine (SVM) was trained. Cell count for each category was used as a feature set for training an SVM. Their proposed method achieved an AUROC value of 0.90 which is significantly better than that of the state-of-the-art method [131] by 0.13.

5.3 MMR Protein Interpretation

The four MMR proteins (MLH1, PMS2, MSH2 and MSH6) exhibit nuclear staining patterns and are assessed for the presence and absence of brown staining produced due to the reaction of the DAB substrate. MMR staining is likely to be observed in any normal cells such as lymphocytes, fibroblasts and normal epithelium. However, the MMR expression is only evaluated in nuclei of the tumour region while comparing their staining to that of normal cells acting as internal controls. Internal controls validated for the presence of staining are termed internal positive controls. In the absence of staining in internal control, MMR protein expression cannot be evaluated and needs repeated IHC tests either on the same or a different tissue block. A clear loss of nuclear staining in tumour cells in the presence of valid internal positive controls would indicate the Loss of protein expression, while a sample with unambiguous staining of tumour nuclei in the presence of valid internal positive controls would be identified as Intact protein expression. A loss of expression of DAB signal in any one of the MMR slides indicates that the patient may have germline mutation causing MMR deficiency, while the presence of DAB staining for all four MMR slides would designate the tumour as pMMR status. In Figure 5.1, visual fields extracted from WSIs of various staining patterns
Figure 5.1: Visual fields taken from three cases with different MMR staining patterns. The tumour region is highlighted with a red contour for the images in the first column while the bottom image contains all tumour. The border colour represents the MMR status where green and red are used for patches extracted from WSIs with intact and loss status respectively. Top row: intact status for all MMR markers; Middle row: intact status for MLH1 and PMS2 while loss status for MSH2 and MSH6; Bottom row: loss of MLH1 and PMS2 while the intact status for MSH6 and PMS2. Notice the nuclear staining of internal positive controls (lymphocytes and stromal cells).

Typically, MMR samples tend to show a strong DAB signal. However, variability in the intensity of DAB staining can be observed from sample to sample. The staining variability in a tissue of the same sample is also not uncommon. For a sample to be given the Intact status, staining of the tumour nuclei should be at least equal to or more than that of the internal positive control. As mentioned earlier, MMR proteins exhibit nuclear staining. However, there are chances of cytoplasmic staining. The possible causes of such unusual occurrences could be issues in IHC testing or a recently reported germline abnormality resulting in fusion of EPCAM and MSH2 [136]. In the presence of both cytoplasmic and nuclear staining, the IHC test is either repeated on the same tissue block or a different tissue block while only a cytoplasmic staining pattern is considered as a Loss of expression.

In a sample with Intact status, it is not uncommon to have a tissue area with the absence of protein staining. There are no standard criteria for an exact cutoff for normal protein expression. There are different propositions
Figure 5.2: A block diagram showing the overall flow for the a) training and b) testing of our 2-step MSI classification pipeline. In this study, we experimented with CK8/18 and H&E as target WSIs.

in the literature for a cutoff of 1%, 5% or 10% [137] while The College of American Pathologists suggests that any amount of positive staining in the tumour nuclei is indicative of Intact status.

5.4 The Proposed Method

A block diagram showing the overall flow for training and inference of the proposed method for prediction of the MSI status is shown in Figure 5.2. There are several pre-processing steps: tumour epithelial segmentation and registration of MMR and target slides, followed by two main modules: MMR prediction and MSI prediction. These pre-processing steps and modules are described in the following sections.
5.4.1 Tumour Epithelial Segmentation

We performed segmentation of the epithelial region in the tumour in two steps: tumour segmentation and epithelial segmentation. The resulting tumour and epithelial mask (binary) were multiplied to obtain a final tumour epithelial mask.

Tumour Segmentation

We trained a classification model using a publicly available colorectal adenocarcinoma dataset released with [138]. This dataset consists of 300 visual fields extracted at 20× magnification of size around 7000×5000 pixels. These visual fields either belong to normal tissue or cancerous tissue. The cancerous visual fields were further classified into two categories: low-grade and high-grade cancer. Of 300 images, normal and low-grade classes comprise 120 images per class while the high-grade category comprises 60 images. We extracted patches of size 224×224 pixels for network training. We trained the Xception model using an Adam optimiser with cross-entropy as an objective function. An example H&E slide with its network generated mask is shown in Figure 5.3. We generated a tumour mask for the CK8/18 image by applying geometrical transformation on the H&E mask. We used transformation parameters which were computed for registering the H&E image to the CK8/18 image. A transformed H&E mask is shown in Figure 5.4.

Figure 5.3: An example of tumour mask overlaid (in green) on the downsampled H&E stained image. On the right side, zoomed version of the regions highlighted in red and blue boxes are shown. The top patch belongs to the normal tissue area with some tumour region at the bottom left while the bottom patch shows the tumour area.
Figure 5.4: Tumour mask for H&E and CK8/18. Tumour masks overlaid on H&E and CK8/18 are shown in the left and right panels, respectively. A tumour mask was originally generated for H&E using a trained CNN model. The tumour mask for CK8/18 was generated by transforming the H&E mask using a transformation matrix computed using the DFBR method proposed in Chapter 4.

**Epithelial Segmentation**

The MMR analysis should be restricted to the epithelial area in cancerous tissue since the presence or absence of positive staining in cancerous epithelial nuclei determines the status (Intact/Loss) of MMR markers. Taking this into consideration, one possible approach could be to perform epithelial nuclei segmentation. To this end, we used a publicly available dataset known as CoNSeP for training a cell segmentation network known as HoVerNet [135]. This dataset consists of 41 visual fields taken from 16 H&E stained colorectal adenocarcinoma WSIs. Each visual field is of size $1000 \times 1000$, extracted at $40 \times$ magnification. We performed stain deconvolution on input images to extract the H channel so that the same network could be used for the H&E and CK8/18 stained images. However, we didn’t use nuclei segmentation masks for our analysis since the trained model didn’t perform well. An example image from CK8/18 slide, its H channel and Hovernet predicted nuclei segmentation mask overlaid on H channel are shown in Figure 5.5. Another possible approach could be segmenting the epithelial layer which is less complicated in comparison to nuclei segmentation in terms of computational time and labelled data creation for training a model.
Figure 5.5: Visual results of nuclei segmentation using HoVerNet. a) CK8/18 image, b) H stain channel extracted from a), and c) nuclei segmentation mask overlaid on b). Blue colour is used for outlining the boundary of epithelial cells. Red blocks are used to highlight the area where epithelial nuclei are missed by the model.

Unlike tumour segmentation, we didn’t transform the H&E mask for generating the CK8/18 mask. Instead, we followed different approaches for CK8/18 and H&E images. This is due to the differences in tissue architecture (mostly visible at higher magnification levels) between H&E and CK8/18 images. We performed epithelial segmentation for H&E images using a UNET-like architecture. We used a publicly available dataset known as CRAG [139, 140]. This dataset consisted of 213 visual fields of size 1500×1500 pixels at 20× magnification, provided in two splits for training and validation. The train and valid sets comprise 173 and 40 visual fields, respectively. We manually annotated the epithelial layer in these images. To improve model performance on the COMET dataset, we also added 23 visual fields of size 2048×2048 from the COMET dataset for training purposes. We performed our experiments using input images at 20× and 10× magnification and didn’t find a significant difference between their average DICE values. To accelerate the mask generation step, we finally used a model trained with images obtained at 10× magnification. We obtained a mean DICE value of 0.8645 on the CRAG validation set. The normal epithelial region was excluded by multiplying the predicted mask with the tumour mask.

Generating an epithelial mask for CK8/18 images was relatively straightforward. CK8/18 are known to have a cytoplasmic and membranous expression in glandular epithelial cells (both normal and cancerous). Considering this, an initial mask was created by a thresholding operation on the DAB channel and then we multiplied this mask with the tumour mask to get a final mask for the tumour epithelial region. The predicted epithelial masks overlaid on the original CK8/18 and H&E stained images are shown in Figure 5.6.
Figure 5.6: Epithelial mask generated by a trained CNN model. The top and bottom rows show overlaid images of CK8/18 and H&E images respectively. Images shown in the third column contain normal epithelium which was excluded using the respective tumour mask.

5.4.2 MSI/MSS Classification

In previous studies, MSI classification is a one-step process where input is the H&E image and output is either MSI or MSS. Our proposed approach makes use of the MMR labels for predicting MSI status and hence we perform MSI prediction in two steps: MMR prediction followed by MSI prediction.

MMR Prediction

There could be at least two different strategies for learning MMR labels, either train separate CNN models for predicting each MMR label individually (as in [141] for ER/PR/HER2 prediction) or train a single multi-output CNN model. In this work, we opt for the latter approach by implementing a multi-headed CNN architecture where each head is responsible for predicting a particular MMR staining expression status. The proposed CNN architecture is shown in Figure 5.7. We employed a partial Xception network as a backbone model while removing the last fully connected layer. The default input size for the Xception model is $299 \times 299$. However, we selected a patch size of $512 \times 512$ pixels to include a bit more context. With an input size of $299 \times 299$, the feature map generated by the last layer before pooling is of dimension $10 \times 10 \times 2048$ while with the $512 \times 512$ input, $16 \times 16 \times 2048$ feature map is generated. To reduce the spatial dimension of the resulting feature map, we added a convolutional layer before the global average pooling layer. After global pooling, the network is divided into four heads, each head comprises three fully connected layers. To prevent the model from overfitting, we added a dropout layer in the first two dense layers with a dropout rate of 0.5.
The ground truth data for the COMET dataset comprises MMR expression status (Intact/Loss) for a WSI. If we follow a weakly supervised strategy for training, then the WSI label could be assigned to each patch extracted from that particular WSI. However, as mentioned earlier, the MMR stained tissue slides with Intact status are likely to have some tumour regions without staining or variable staining. Keeping this into consideration, we adopt stain co-localisation for generating labels locally, rather than globally. We term this as a ‘localised’ setting for generating training data. Using this setting, we would get four binary labels against each input patch, where each label corresponds to an individual MMR protein. To this end, we first register the MMR slides to the target slide (CK8/18 or H&E) using our method proposed in Chapter 4. We then extract a patch from a target slide at 20× magnification using a tumour epithelial mask and the corresponding patches from the registered MMR slides. If the MMR expression label for a WSI is ‘Loss’ then we assign the ‘Loss’ label to every patch of that WSI. However, if the MMR expression label is Intact then we first compute the DAB stain intensity of a patch using the H-Score (as described below) and then perform thresholding to decide if the patch has Loss or Intact status. A pictorial demonstration of generating labelled data is provided in Figure 5.8. To compute the H-score, cellular staining intensity is first divided into four categories: strong (3+), moderate(2+), weak(1+) and no staining/negative(0). The summation of the weighted percentage of these intensity categories would compute the final H-Score value. We performed the following steps for computing the H-Score:

1. Deconvolve the IHC patch into H and DAB channels;
2. Threshold H channel to detect nuclei;
3. Using nuclei mask, we apply thresholding to each nuclei pixel in the DAB
channel such that they fall into one of the four categories: strong ($\geq 0$ \& $\leq 60$), moderate ($\geq 61$ \& $\leq 150$), weak ($\geq 151$ \& $\leq 200$) and negative ($\geq 201$ \& $\leq 255$). Thresholding is performed after taking the complement of the DAB channel;

4. Finally, we compute the H-Score using the following formula [142, 143]:

$$
H\text{-Score} = 3 \times (\% \text{ of nuclei stained at intensity category '3+') + 2 \times (\% \text{ of nuclei stained at intensity category '2+') + 1 \times (\% \text{ of nuclei stained at intensity category '1+') +
$$

5. The H-Score ranges between 0 and 300. On the basis of a discriminatory threshold, the patch status was considered either Intact or Loss. We selected a threshold value of 130.

We then train our multi-headed network using target input patches generated using our localised setting. We trained the model using an Adam optimiser with binary cross-entropy as an objective function for each MMR label. We performed stain augmentation, rotation, flipping and elastic distortion for augmenting our training patches. Once the network is trained, we proceed to our second step for MSI prediction where we employ the trained model for generating MMR probability maps. We then extract features from these maps and utilise them for classifying each case as MSI or MSS.

**MSI Prediction**

Using the trained MMR model, we generate probability maps for all the train and test cases. For any given case, we get four probability maps, one per MMR biomarker. We then compute features for each MMR protein which is the histogram of probabilities of tumour epithelial patches. We use two histograms: one for protein Intact status and the other for protein Loss status. We selected the number of histogram bins parameter as 15 and normalised the histogram values. Finally, we input these features to an SVM classifier for WSI level MSI prediction. A diagrammatic illustration of the MSI prediction step is shown in Figure 5.9. Note that MMR slides are only used during the training phase for labelled data generation. During inference we only use target slides for MMR and MSI prediction, hence ceasing the time burden of the pre-processing step of registration.

**5.5 Dataset and Performance Measures**

We used the COMET dataset comprising of 86 cancerous cases for our experiments on MSI prediction, as discussed in detail in Section 3.4.1. For WSI-level
Figure 5.8: Co-localised MMR Prediction using CK8/18 as a target image. a) Downsampled WSIs of MMR IHC sections registered to CK8/18 stained WSI. b) An example input image used for training a CNN. Its ground truth label is defined by thresholding H-score values computed for the corresponding patches from MMR slides. The corresponding MMR patches are shown in c). The colour of the borderline in c) shows the expression status for the given patch. Green represents that the patch is positively stained with DAB while red shows the absence of DAB stain. Patches showed in b) and c) are highlighted in red circles in a).
Figure 5.9: Overview of the proposed MSI prediction step. The MMR probability heatmaps and corresponding normalised histogram of probabilities are shown for both MSS and MSI cases.

Table 5.1: MMR ground truth labels for the COMET dataset.

<table>
<thead>
<tr>
<th></th>
<th>MLH1</th>
<th>MSH2</th>
<th>MSH6</th>
<th>PMS2</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSS</td>
<td>Intact</td>
<td>Intact</td>
<td>Intact</td>
<td>Intact</td>
<td>56</td>
</tr>
<tr>
<td>MSI</td>
<td>Loss</td>
<td>Loss</td>
<td>Loss</td>
<td>Loss</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Intact</td>
<td>Loss</td>
<td>Loss</td>
<td>Intact</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Loss</td>
<td>Intact</td>
<td>Intact</td>
<td>Loss</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Intact</td>
<td>Loss</td>
<td>Intact</td>
<td>Intact</td>
<td>1</td>
</tr>
</tbody>
</table>

MMR labels, we provided MMR tissue images along with their corresponding H&E to a pathologist. Of the 72 tumours out of 86, in which MMR analysis was possible, we obtained the expression status for each MMR slide. We then categorise these 72 cases into MSI and MSS using their MMR labels. Cases for which all the MMR slides were given Intact status were classified as MSS while any case with at least one MMR slide with Loss status was categorised as MSI. In total, we obtained 16 (22.2%) MSI and 56 (77.8%) MSS cases. The count of cases with different MMR and MSI ground truth labels are presented in Table 5.1. We assess the performance of our method by 5-fold cross-validation while keeping the same train and test set for CNN (MMR prediction) and SVM (MSI prediction) training and testing. We evaluated patch-level MMR prediction using DICE overlap and WSI-level MSI prediction using AUROC. To investigate the generalisability of our trained CNN model, we predicted MMR status in the patient cohorts from The Cancer Genome Atlas (TCGA).
Target | MLH1 | MSH2 | MSH6 | PMS2  
--- | --- | --- | --- | ---  
CK8/18 | 0.720 ± 0.045 | 0.829 ± 0.036 | 0.885 ± 0.028 | 0.721 ± 0.053  
H&E | 0.727 ± 0.040 | 0.797 ± 0.032 | 0.870 ± 0.019 | 0.732 ± 0.044

Table 5.2: Average DICE values computed for each MMR marker across 5-fold cross-validation.

5.6 Experimental Results

We carried out two sets of experiments: one with a CK8/18 slide as a target image and the other with a H&E slide as a target image. Below, we present our results of MMR and MSI classification steps for both CK8/18 and H&E slide images.

5.6.1 MMR Prediction

The MMR probability maps for Intact status overlaid on the original MMR tissue images are shown in Figure 5.10, along with the zoomed-in regions. We first computed the DICE overlap ratio to evaluate the performance of our multi-headed CNN model for MMR prediction. To this end, we generated the ground truth H-Score maps by processing MMR slides at 20× magnification in a sliding window fashion. The non-epithelial region was excluded while computing the H Score map. Thresholding was then performed to convert the H-Score maps to binary maps where the foreground represents a region with Intact status while the background represents a non-tissue region or a region with Loss status. The network-generated MMR probability maps were also converted to binary maps using a threshold value of 0.5. The DICE index was then computed between the thresholded H-Score and the network predicted maps. The DICE values for each MMR biomarker computed for all 5-fold cross-validation splits are shown in Figure 5.11 and Table 5.2.

We also performed a baseline experiment by directly utilising the MMR prediction maps for MSI classification, without training an SVM model. Thresholded probability maps were used for deciding the Intact/Loss status for MMR proteins and it was based on the number of patches with Intact/Loss status. The Intact status was assigned if the number of Intact patches was more than the number of patches with Loss status. Finally, the MSI status was identified using these MMR labels. We obtained AUC of 0.8869 ± 0.150 and 0.8500 ± 0.124 for CK8/18 and H&E respectively, across a 5-fold cross-validation data split. Figure 5.12 shows the results of MMR status prediction and its correlation with the pathologists’ MMR labels. On using MMR labels for CK8/18 as a target image, we were able to identify 98% and 75% of MSS and MSI cases, respectively. While for H&E as a target image, we correctly classified 95% of MSS and 56% of MSI cases, respectively.
Figure 5.10: MMR prediction against CK8/18. a-d) MMR stained slides: MLH1, MSH2, MSH6 and PMS2, e-f) respective probability heatmaps (generated by our multi-headed network) overlaid on registered MMR stained WSIs and the final predicted WSI label is MSI, i-l) zoomed in patches highlighted on a-h with small rectangles. The ground truth MMR expression status for a-d is Loss, Intact, Intact and Loss.
Figure 5.11: Evaluation of our MMR prediction step using the DICE coefficient. The overlap is computed between the network generated MMR binary maps and ground truth H-Score maps after thresholding.

5.6.2 MSI Prediction

We trained SVM models using histogram features with different kernel functions: linear, polynomial and RBF. We utilised our multi-headed network trained using two different settings for extracting histogram values. The localised setting is the one discussed in Section 5.4.2 while for the unlocalised setting, slide-level MMR labels were assigned to all the patches rather than performing stain co-localisation analysis for generating labels. The average AUCs are reported in Table 5.3. The SVM with RBF kernel has been shown to outperform other kernel functions for both CK8/18 and H&E irrespective of any training setting. Our results suggest that the histogram values generated in the localised setting are more effective in classifying cases into MSI/MSS status as compared to those generated in an unlocalised setting. ROC curves for both CK8/18 and
Figure 5.12: Scatter plot showing the WSI-level MMR prediction results using our baseline experiment in comparison to that of the pathologist’s assignment. The corresponding confusion matrix is shown below each scatter plot. Any MMR slide with Intact status is given a value of 1 while a slide with Loss status is given 0. This plot was generated by taking the sum of these MMR values. The size of the circular markers is relative to the number of cases with matching pathologist and algorithm generated labels.
Table 5.3: Average AUROC values for MSI prediction using 5-fold cross-validation obtained with different kernel functions in SVM.

<table>
<thead>
<tr>
<th></th>
<th>Linear</th>
<th>Polynomial</th>
<th>RBF</th>
</tr>
</thead>
<tbody>
<tr>
<td>CK8/18</td>
<td>Localised</td>
<td>0.9596 ± 0.048</td>
<td>0.9444 ± 0.062</td>
</tr>
<tr>
<td></td>
<td>Unlocalised</td>
<td>0.9051 ± 0.124</td>
<td>0.9414 ± 0.055</td>
</tr>
<tr>
<td>H&amp;E</td>
<td>Localised</td>
<td>0.8152 ± 0.176</td>
<td>0.8455 ± 0.149</td>
</tr>
<tr>
<td></td>
<td>Unlocalised</td>
<td>0.7798 ± 0.179</td>
<td>0.7859 ± 0.175</td>
</tr>
</tbody>
</table>

H&E based predictions for 5-fold cross-validation are shown in Figure 5.13.

Figure 5.13: MSI Prediction on CK8/18 and H&E images using localised and unlocalised settings. AUROC plots of 5-fold cross-validation along with average AUROC is shown. For each curve, x- and y-axis represent false positive and true positive rates.

5.6.3 Comparative Analysis

MSI cases can have different patterns of MMR expression, as observed in our dataset as well (see Table 5.1). Hence, in this regard, MMR labels can be considered as fine-grained labels of the coarse-grain MSS/MSI labelling. To the best of our knowledge, it is the first time that MMR labels are utilised for microsatellite (MS) classification. It is equally correct assigning MSI status to a case having an MMR pattern where at least one of the MMR proteins has a Loss status. Since our main task at hand in this study was to classify
coarse-grained labels (MSS vs MSI), a question can arise: why not train a CNN using coarse-grain labels as it has been done in previous studies on MSI classification [127, 130–134]. In other words, is it beneficial to first train a CNN model with fine-grain labels (MMR labels) and then map them to coarse-grain labels (MSI/MSS) in a second phase?

Chen et al. [144] have shown that the CNN trained on natural images with fine-grained labels achieves better performance as compared to the network trained with coarse-grained labelled images of the same dataset. To study this with our dataset, we conducted another set of experiments, in which CNN was directly trained on MSI/MSS labelled patches instead of MMR labels. We trained the Xception network using Adam optimiser with binary cross-entropy as an objective function. The samples with a high proportion of MSI predicted patches to the total sum of MSI and MSS patches were labelled as MSI cases, as in [131]. We obtained AUC of $0.8646 \pm 0.1152$ and $0.8152 \pm 0.1019$ using the CK8/18 and H&E stained slides, respectively. In addition, similar to our approach discussed in 5.4.2, we trained an SVM using two normalised histograms of probabilities: one for MSS and the other for MSI and obtained AUC of $0.8465 \pm 0.1391$ and $0.8152 \pm 0.1166$ with CK8/18 and H&E stained slides, respectively. These results demonstrate that MSI classification can be improved by training CNN with fine-grain labels (MMR labels).

5.6.4 TCGA Cohort

We experimented with colon and rectal cancer cohort (TCGA-CRC-DX) which was used in previous studies [130, 131, 133] on MSI prediction. The TCGA-CRC-DX cohort comprises a total of 428 WSIs, obtained from 423 patients. The ground truth labels of the cohort contain only MSI labels, not MMR expression status. To this end, we retrained our multi-headed network with all the H&E stained WSIs in the COMET dataset, followed by MMR prediction on the TCGA dataset. The training was performed in two separate settings: stain augmentation and stain normalisation.

On visualising MMR prediction heatmaps overlaid on the H&E images, we mainly observed two patterns of MMR expression: pattern where most of the tumour regions are predicted as Intact (or positive) for all the MMR markers (see Figures 5.15 and 5.16) and patterns where most of the tumour regions are predicted as Intact for MSH2 and MSH6 markers and Loss (or negative) for MLH1 and PMS2 markers (see Figures 5.14 and 5.17). We found that the trained model didn’t generalise well for the TCGA dataset. Unlike our observation for the COMET cases, the aforementioned patterns were observed in both MSI and MSS cases, hence MMR heatmaps could not be used to extract discriminatory features for the TCGA dataset. Few example images in MSI
and MSS subgroups with the overlaid predicted MMR heatmaps are shown in Figures 5.14 to 5.17. In Figures 5.14 and 5.16, we observed expected MMR expression for categorising this case as MSI and MSS, respectively, whereas in Figures 5.15 and 5.17, we observed unexpected MMR expression for categorising this case as MSI and MSS, respectively. We observed similar findings with both models, irrespective of their training strategies (stain augmentation or stain normalisation).

In computational pathology, model generalisation is a major challenge due to the domain shift. It is particularly prevalent when the model is trained on a small dataset, obtained from a single site with less diversity in terms of sample preparation and image acquisition. The problem of generalisation in our experiments is due to the same reason, the domain shift. Below, we list some possible elements contributing to the domain shift:

- Tissue samples for COMET were collected from a single site and all WSIs were acquired using the Omnyx VL120 scanner while following the same acquisition protocols. The TCGA slides are collected from many sites across the USA. In addition to the multi-site aspect, the heterogeneity can also be characterised by the scanner modalities, manufacturers and acquisition protocols [145].

- The COMET and TCGA slides differ in terms of the stage of cancer. COMET slides were taken from the cancerous region described as either stage 3 or 4 while TCGA samples were taken from stage 1-4 cancers.

- We have looked at the distribution of histological type by MSI status in mucinous and non-mucinous cancer for both the COMET and TCGA cohort, as shown in Figure 5.18. The number of mucinous samples with both MSI and MSS status is significantly low for both datasets. Since our CNN model is trained on the COMET dataset, a total of 4 and 2 mucinous samples for MSI and MSS respectively are not adequate for learning discriminatory features for classification.

5.7 Summary/Discussion

In this chapter, we have presented a novel method for MSI classification in a two-stage process. The main focus is to perform MMR biomarker estimation from a single target image (CK8/18 or H&E) with the end goal of identifying cases with MSI status. To this end, we utilised registered images of the target and MMR tissue slides for identifying tissue regions with stain Intact/Loss status in a localised manner, rather than assigning a WSI label to every patch. Even with imperfections in MMR prediction maps, SVM was able to identify
Figure 5.14: Example images from the TCGA cohort labelled as MSI. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has performed well for these images.
Figure 5.15: Example images from the TCGA cohort labelled as MSI. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has not performed well for these images.
Figure 5.16: Example images from the TCGA cohort labelled as MSS. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has performed well for these images.
Figure 5.17: Example images from the TCGA cohort labelled as MSS. Our MMR model assigned probabilities of being positive (Intact) or negative (Loss) are used to generate a heatmap overlay for the tumour region only. In the heatmap, red and blue colours correspond to the Intact and Loss class, respectively. The network has not performed well for these images.
samples with MSI status. Our comparative analysis demonstrated improved performance if the CNN model is trained in a localised setting. Similar to the findings in previous literature, CNN trained with fine-grain MMR classes outperformed the CNN trained with coarse-grain MSI/MSS classes.

In clinical practice, identification of dMMR cases either using PCR or IHC testing incurs cost since it requires more tissue sections in addition to H&E stained section. Due to its high cost, not all patients are investigated which may lead to an inappropriate treatment plan for the patient. Patients with Lynch syndrome can potentially be imposed to chemotherapy rather than treated with immunotherapy, due to the lack of testing. The introduction of computational tools may assist the pathologists to overcome these issues and improve diagnostic performance and patient outcomes. The results of biomarker status prediction using computational methods such as those presented in our study are encouraging and can open routes for their applications in clinical workflows after further improvement and validation at a larger scale.
Chapter 6

Conclusions and Future Directions

In this thesis, we presented a set of computational approaches that can help identify patients for a particular treatment or those with a high risk of malignancy. First, we presented our work on bladder cancer for identifying high risk cases which require more aggressive follow-up and a different clinical management approach as compared to those with a low risk of malignancy. Second, we presented our work on registering multi-IHC WSIs which serves as an essential pre-processing step for our multi-slide analysis work. We presented two different methods for registration, one using hand-crafted features and the other using data-driven features. Third, we presented our work on multi-IHC analysis for MSI prediction from the H&E and CK818 stained images while using registered MMR images.

In the following sections, we present a summary of each chapter while discussing the main contributions, limitations of the proposed approaches and possible future extensions to improving the proposed methods.

6.1 Identifying Patients with High Risk for Bladder Cancer

The main goal of this study is to investigate an automated alternative to risk stratification of urine cytology slides since the status quo based on subjective visual analysis has large inter- and intra-observer variability and is far from perfect. We first iteratively obtained cell-level annotations from expert pathologists who followed the TPS criteria for labelling. To balance the cellular distribution in the training set, we performed over-sampling using an existing method known as SMOTE. With the balanced training set, we trained an Xception network for cell classification. During WSI processing, we performed two steps before classifying cells. We first performed ROI extraction from WSI
to confine our processing to the region with cellular content and in the second step, we identified candidate cells for further processing using the trained Xception model, for further processing. For both these steps, we estimated optimal threshold values using the Otsu algorithm. We then utilized a cell count based approach for identifying cases with a high risk for bladder malignancy. We also investigated the inter-observer agreement at the WSI level and intra-observer variability at the cell level and rationales for the discrepancies. Lastly, we investigated the cytopathology based risk category and our digital risk labelling in terms of its correlation with the ‘gold standard’ histopathology based diagnosis.

We employed ROC analysis as an evaluation metric for our cell and WSI level classification. For cell classification, We considered atypical and malignant cells as positive classes and the remaining classes as negative for plotting the ROC curves. We obtained the area under the ROC curve value of 0.99 for the validation set. For WSI classification, we used the count of malignant cells (M) and the total count of atypical and malignant cells (A+M). We obtained an average AUC of 0.81 and 0.83 using M and A+M, respectively. We also found the digital risk to be a better predictor of the histopathology based diagnosis.

One possible future direction is the phenotyping of the atypical and malignant cells based on the morphological features of the detected cells and nuclei. The extension of this work could focus on analysing these morphological features in relation to cytopathology based and histopathology based diagnoses.

The proposed method for automated risk stratification of urine cytology slides has demonstrated the promise of deep learning based risk scoring for bladder cancer. However, the data used in this study is obtained from a single centre that may have introduced a bias into the machine learning model. Therefore, we will need to conduct large-scale multi-centric trials for establishing the efficacy of the proposed method. In addition, cell-level annotations, particularly for the diagnostic classes, were mainly obtained from a single pathologist. We observed intra-observer variability, potentially undermining the accuracy of our model which could perform better if the agreement rate was high. We also observed disagreement at the WSI level. Currently, clinical assessment of cytology slides is manual and it is because the pathologists frequently need to focus on different planes to view all the cells in a cell cluster. The WSI is captured in a single plane and one can argue that introducing the z-stacking feature in WSIs can potentially help with an accurate assessment. However, it comes at the cost of a much larger image file size and longer scanning and analysis times.
6.2 Registration of multi-stained WSIs

For registering multi-stain WSIs, we first extended an existing maximally stable extremal regions (MSER) feature based method by performing the alignment in a multi-scale manner. We employed feature matching of maximally stable extremal regions at three consecutive magnification levels (0.3125×, 0.625× and 1.25×) to refine the registration at the immediately next magnification levels. The multi-scale approach improved the registration accuracy at the cost of increased processing time. It is vital to perform registration in a reasonable time such that it does not hinder the downstream analysis. With MSER, the processing time for registering a pair of images depends on the number of valid MSERs; hence it varies with the input image pair. It can be impractical to perform registration in the presence of a large number of MSERs. Therefore, to improve the processing time, we proposed a deep feature based registration (DFBR) approach that utilizes data-driven features for alignment. We extracted these features using a pre-trained VGG model. Unlike MSER, we obtained a fixed number of features for each image using a CNN; hence the processing time remained the same for any pair of images. This does not only make the feature matching step faster as compared to that with MSER features, but it also improved the registration accuracy. We used matching feature points for finding the rigid transformation and to deal with the non-linear deformities, we employed an existing non-linear registration method which won the ANHIR challenge.

Our DFBR approach is shown to outperform both single and multi-scale approaches using MSER features. The set of images that we used for evaluation comprises some image pairs with significant morphological changes. We showed that our DFBR approach was able to align those challenging image pairs better than multi-scale MSER and case-wide approaches. We obtained average median rTRE 0.0955, 0.0358, 0.0380 and 0.0063 with single-scale MSER, multi-scale MSER, case-wide and DFBR approaches respectively. Using the ANHIR dataset, we obtained comparable results to the challenge winning team. The current approach is implemented to find the optimal control points for estimating rigid transformation. While the proposed feature matching approach can be adapted for non-rigid alignment.

We implemented two interfaces to visualise the original reference and transformed source images side-by-side. The first interface applies the rigid transformation to the WSI as we zoom and pan around the image. To do so, the global transformation matrix initially computed for a WSI is transformed such that it can be applied locally to the tiles displayed on the screen. The second interface can be used to visualise the registered images where the source image has been transformed and converted to the pyramidal format. Currently,
the registration process and visualisation are carried out separately and it may be desirable for there to be an interface that takes two images as an input, computes the transformation and then displays the registered images. In the current version of the interface, the user can only visualise image pairs side-by-side. In the extended version, it would be useful to have a curtain viewer [146] to enable the user to visualise overlapping registered images.

In the visualisation interface, we added a function to enable the user to fix an offset when observed during the visualisation. Currently, the added functionality can only compute a translation offset based on the current FOV. However, the reference and registered moving FOVs are likely to have rotational offset as well. Therefore, it would be desirable to tackle both translational and rotational offsets in the visualisation interface. Additionally, the interface may also have an option to perform registration manually by selecting corresponding points in two images. This could be used when the automatic method fails to give optimal registration.

The multi-IHC dataset that we used for registration and MSI prediction comprises sections stained with IHC biomarkers other than MMR proteins. Once the whole stack of tissue sections is registered, one can analyze these stains for more in-depth analysis. The possible analysis could be scoring of Ki67 and p53 and their association (if any) with the MMR expression patterns or investigating the local expression patterns for predicting clinical outcomes.

6.3 Computerised Assessment of Mismatch Repair Status for MSI Prediction

Following the registration step, we performed a multi-slide analysis to identify colorectal cancer patients with MSI tumour. In previous studies, CNN models are trained using MSI/MSS labelling. In our work, we trained our model with a bit more information comprising various MMR staining patterns. We performed MSI/MSS classification in a two-step process. In the first step, we generated a probability map representing the chances of the WSI region having an Intact status. To this end, we trained a four headed CNN network, each head responsible for outputting the probability of Intact status for the corresponding MMR biomarker. For CNN training, we extracted patches from the target image and obtained the ground truth labels using the stain colocalisation approach. The binary labels are the thresholded H-score values computed for corresponding patches from MMR slides. In the second step, we used these probability maps for deciding MSI status with the help of an SVM classifier.

We compared the results of MSI prediction using our localised MMR
prediction with that of unlocalised MMR prediction. For the unlocalised version of the experiments, we assign the MMR label at the WSI level to all the WSI patches irrespective of the actual DAB staining in the MMR sections. We obtained average AUC values of 0.96 and 0.94 with localized and unlocalized approaches, respectively using the CK818 image. When using the H&E image, we obtained average AUC values of 0.92 and 0.83 with localised and unlocalised approaches, respectively. In another experiment, we performed MSI prediction using MSI labels instead of MMR labelling and obtained average AUC values of 0.86 and 0.82 with CK818 and H&E, respectively.

The visual assessment of MMR markers using IHC has been shown to suffer from observer variability [147, 148]. In our work, we obtained MMR expression status from a single pathologist that may have introduced a bias and we consider it as a limitation of this work. The DAB staining for MMR markers is only relevant in tumour epithelial regions and to confine our analysis, we generated a tumour epithelial mask. In the current work, we trained two different networks, one to generate the mask for the tumour region and the other for generating the epithelial mask. This could be accomplished by training a single network to accelerate the processing time.

In this work, we have assessed the MMR biomarkers in a binary manner per the guidelines for MMR IHC assessment [149]. This could be expanded upon by considering the DAB staining in a continuous spectrum, as there is a possibility that we may be able to find an association between the MMR continuous expression profiles based, for instance, on H-score for better prediction of MSI status and response to therapy. We compared our proposed approach with one previous study [131] and in future, we will evaluate it against a recently proposed method [130] outperforming [131].

We also investigated the performance of our model for MMR prediction on the TCGA dataset for colorectal cancer using H&E images. We trained the model using all the slides in the COMET dataset and then tested it on the TCGA dataset. On visualising predicted heatmaps, we found that model didn’t generalise well to the TCGA dataset. In addition to the small training dataset, we suspected domain shift as a main contributing factor to the generalisability problem. To investigate it, we explored the differences between the COMET and TCGA datasets that are likely to cause the domain shift.

6.4 Concluding Remarks

The work presented in this thesis demonstrates that deep learning-based algorithms often promise to identify people who are at high risk of bladder cancer and to select colorectal cancer patients for immunotherapy. We achieved encouraging results using a dataset obtained from a single centre. However,
the lack of sufficient training data for training data-hungry models is often the hurdle in increasing the generalisability of the proposed methods. In addition, the models should be trained and validated on large multi-centric cohorts for better assessment and evaluation. For clinical applications, these methods need to be evaluated on independent cohorts of patients with wide diversity in terms of demographics, tissue preparation and staining to establish the efficacy of the proposed methods in the true sense.
Bibliography


[23] S. Thomas, “Pathologist versus artificial pathologist: What do we really want (need) from machine learning,”


