DETECTION OF WESTERN BLOT PROTEIN BANDS USING TRANSFER LEARNING IN THE ASSESSMENT OF RISK OF GASTRIC AND OESOPHAGEAL CANCER

Girmaw Abebe Tadesse[1,2], **Daniel Chapman**[2], **Ling Yang**[2], **Pang Yao**[2], **Iona Millwood**[2], **Zhengming Chen**[2], **Tingting Zhu**[2]

[1]IBM Research - Africa, Nairobi, Kenya [2] University of Oxford, UK

ABSTRACT

We present a transfer learning technique to encode visual features from an immunoblot assay for the detection of immuno-reactive protein bands to understand the association between H. pylori infection and gastric and oesophageal cancer. A CE-marked immunoblot method (HelicoBlot 2.0) was used to analyse 1500 human serum samples, test strips were scanned and the images segmented to enable a machine-learning algorithm to be developed which could identify protein bands on the strips. A model has been developed to detect protein bands with a performance of 95% AUROC. This novel approach to protein band detection reduces time spend by laboratory experts to interpret results, reduces ambiguity in band classification, and can be applied to large-scale epidemiological studies investigating the relationship between infectious disease and risk of cancer.

1 INTRODUCTION

The associations between H. pylori infection and gastric and oesophageal cancer are either inconsistent or still poorly understood. As part of the China Kadoorie Biobank (CKB) Cancer Research UK (CRUK) infection and gastro-intestinal cancer project, H. pylori infection status among 500 cases each of cardia and non-cardia gastric cancer and 500 randomly selected sub-cohort individuals was determined using a CE-marked immunoblot method – HelicoBlot 2.0 (MP Biomedicals). The HelicoBlot assay provides qualitative detection of IgG antibodies specific to proteins of H. pylori, including antigens associated with pathology such as CagA and VacA (Parsonnet et al., 1997; McClain et al., 2017)

There are over 18 immuno-reactive protein bands present on HelicoBlot test strips, however, only 7 of these are recognised by the manufacturer as relevant to diagnose H. pylori infection. These proteins are identified by their molecular weight and named p19.5, p30, p35, p37, p89 (VacA), p116 (CagA), and CIM (current infection marker). Prior to testing CKB study samples, the HelicoBlot assay underwent validation to assess its analytical performance. The assay met the desired performance criteria, and it was decided that each strip would be interpreted by two independent operators when analysing CKB samples. Furthermore, the presence of diagnostic bands would be recorded against molecular weight and scored 0 = negative (no band seen), 1 = ambiguous result, 2= clear positive (band present). Where observers disagree the HelicoBlot was scored by consensus. If a consensus was not reached, sample analysis needed to be repeated. Thus, identifying and classifying protein bands on test strips was subjective and time consuming.

To this end, we propose a machine learning algorithm in order to reduce interpretation subjectivity and improve sample throughput for future HelicoBlot analysis projects. Machine learning provides a promising potential in automating cumbersome processes across different applications domains via learning from experience (Goodfellow et al., 2016). To investigate the ability of machine learning to interpret HelicoBlot strips, a protocol was established to scan all test strips from the CKB analysis at high resolution.

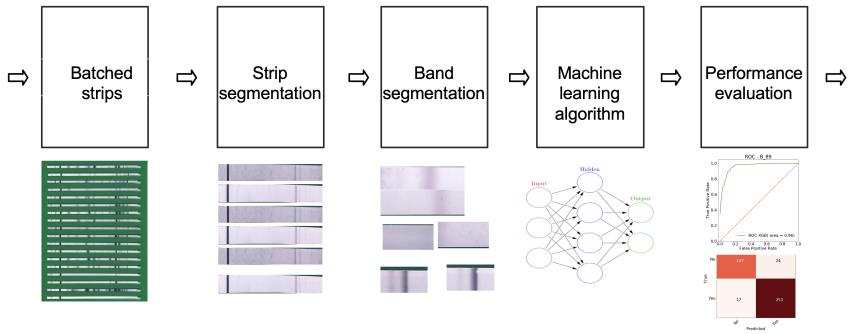

Figure 1: Proposed overview of machine learning-based protein band detection.

2 PROPOSED APPROACH

Batched strips and segmentation High resolution scanning of strips was performed, and eighteen strips were batched together for effective resource utilisation. A bespoke stencil was used to scan a batch of strips (n=18) at the same time (See Fig. 1). Each strip in a batch represent a sample for the machine learning algorithm. Thus, an image processing algorithm is developed to segment each strips separately, followed by the segmentation of the seven protein bands recognised by the HelicoBlot manufacturer from each strip. To do so, we asked a domain expert to annotate the start and end pixels of protein bands in 5 strips. Then we employ the minimum and maximum of start and end pixels, respectively, for each protein band across the 5 strips in order to maximise the existence of a band in a segmented strip.

Machine learning algorithm The machine learning approach is applied to train and infer the presence of a band in each segment. Among the segmented samples, we only used those coming from either 0 or 2 consensus among the annotators. We use existing computer vision network, GoogleNet (Szegedy et al., 2015), in order to extract features from each segment via transfer learning approach (Pan & Yang, 2009). This results in the extraction of 2048-D feature space from the penultimate hidden layer of GoogleNet. Then XGBoost classifier is trained to predict the presence of a protein band using a 70% train and 30% test of stratified cross-validation.

Performance evaluation Area under Receiver Operating Characteristics (AUROC), Sensitivity and Specificity metrics are employed evaluate the performance of the proposed protein-band detection approach. The results are shown in Table 1, and encouraging performance is achieved across all the segments.

	Protein bands							
	CIM	p19.5	p30	p35	p37	p89	p116	**Average**
AUROC (%)	99	90	96	92	92	96	97	**95**
Sensitivity (%)	95	58	93	72	77	94	98	**84**
Specificity (%)	98	95	94	94	92	86	91	**93**

Table 1: Prediction performance of the proposed approach in detecting the different protein bands from the scanned strips of HelicoBlot assays, validating using XGBoost classifier.

3 CONCLUSIONS

A machine learning approach is shown to facilitate the detection of immuno-reactive protein bands present on HelicoBlot test strips. To do so, we employ transfer learning technique by using an existing computer vision network in order to extract features for the XGBoost-based classifier. The encouraging performance achieved in detecting protein bands would help address the ambiguity associated with manual labelling of these strips, and hence facilitates the understanding of H. pylori infection and gastric and oesophageal cancer.

REFERENCES

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Mark S McClain, Amber C Beckett, and Timothy L Cover. Helicobacter pylori vacuolating toxin and gastric cancer. *Toxins*, 9(10):316, 2017.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

J Parsonnet, GD Friedman, N Orentreich, and H Vogelman. Risk for gastric cancer in people with caga positive or caga negative helicobacter pylori infection. *Gut*, 40(3):297–301, 1997.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proc. of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.