Ensemble of Boosted Decision Stumps for Detection of β Cells

*Abstract*— Diabetes has been linked with a reduction in the functional ß cell mass. Thus, to cure it or to stop its progress, this cell mass must be restored. A count of the cell mass of such kind requires determination of the number of nuclei in the insulin stained area. Such a count consumes significant time if done by skilled experts. Also, this count result might vary if carried out by different experts; hence, the results become subjective. In this paper, we propose an approach based on the cascading of weak learners such as decision stumps which are boosted in nature. The learners gather information from a provided dataset and form a series of layers which go on to analyze from the coarser to the finer aspects of the cell image. This method gives significantly better experimental results when compared to the results produced by human experts as well as that produced by the best reported results till now.

*Index Terms*— Beta cell mass, Cell nuclei detection, Weak Learners, Boosted Decision Stumps

# INTRODUCTION

The control of the level of glucose in the human blood is of paramount importance. This level if exceeds a critical value must be brought down. This task is accomplished in the human body by a hormone called insulin which is produced by the β cells present in the pancreatic islets. β cells react to spikes in blood glucose concentrations and release insulin while simultaneously producing more. Thus, more or less all major types of Diabetes involve a reduction in the β cell count which in turn hampers their body’s insulin production capability. As a result much research activity has been carried out to invent therapeutical strategies or genetic manipulations that freezes the loss of β cell mass or restores it to cure the disease[1]. Post-mortem analyses of β and islet cell mass are increasingly undertaken in patients as it is not currently possible to measure changes in β cell or even islet mass during life due to lack of existence of sufficiently sensitive or specific imaging tools.

Much significance lies with the accurate post-mortem analysis of beta cells. Estimation requires Immunohistochemical (IHC) staining of the pancreatic tissue sections. Insulin reacts with this stain and produces an immunoreactive area in the tissue section. The area of this section divided by the total tissue section area is a measure of the beta cell mass. To obtain the mass, this ratio is multiplied by the total pancreas weight. The beta cell count however, requires a count of the number of nuclei within that stained insulin area. This counting procedure is very labour intensive and is observer biased. The process of image capturing combined with the long and tedious analyses of the images thus lead experimentalists to consume months for a single study.

The idea of making this entire cell count process computerized has been taken up several times and as such we do have algorithms for this. The very commonly used algorithms for such cell detection are intensity based such as thresholding[2] and clustering[3,4]. However, the issue with using intensity based algorithms is that such algorithms only provide a satisfactory result when there is significant contrast between the background and the foreground(the cells, in this case). Active contours to segment out the boundary of the cells has been one of the methods explored vastly[3-7]. It iteratively modifies the contour based on some criteria. However it depends on the seed points and can result into oversegmentation. Bamford *et al*.[5] have showed that dual-active contours overcome such shortcomings. The intensity profiling of an image as a topological surface gives way to another method commonly known as watershed. It highly depends on the detection of local minima for segmentation of cells and in noisy and textured images having several irrelevant local minima, such an algorithm does not perform well. If used on a gray image without processing then that again leads to oversegmentation. Initially determining markers for the region of interest makes the process semi-automated and is probably the best method. Other approaches include development of a statistical model[8] and morphological operations[9] for detection of cell nuclei. Kuse *et al.*[17] suggest another method for detection of cell nuclei based on the idea of isotropic phase symmetry which is the state of art and we shall be aiming to improve on that. The algorithm works in the frequency domain and hence noise, contrast, illumination etc. are invariant. It quantifies symmetry about a pixel irrespective of direction(s) where symmetry is found. When applied on Hematoxylin-positive parts of a histology image, this measure can be used to detect cells.

In this paper we propose a novel method for detection of cell nuclei. The main idea behind it is in iteratively training decision stumps using Adaboost to detect all the nuclei and subsequently detect the stained region using color deconvolution to find out the count of β cells.

The proposed algorithm using Ensemble of Boosted Decision Stumps(EoBDS) is depicted as a schematic diagram in Figure 1 and has been discussed in detail below. A transgenic plns-c-MycERTAM mouse model[10] was employed in this study.

# Training A Cascade Object Detector

The object detection framework is a cascade classifer that consists of stages, where each stage is an ensemble of weak learners which are simple classifiers called decision stumps. Such a framework for face recognition has been previously shown to give a considerably good performance by Viola *et al*.[11] The object detector classifies images based on the value of simple features and not the pixels directly. Such a feature based system operates much faster than a pixel-based one. The simple features used by the detector are two-rectangle, three-rectangle and four-rectangle feature which are quite similar to Haar basis functions which have been used by Papageorgiou *et al.*[12] Rectangle features can be computed very rapidly using an intermediate representation for the image commonly known as integral image.

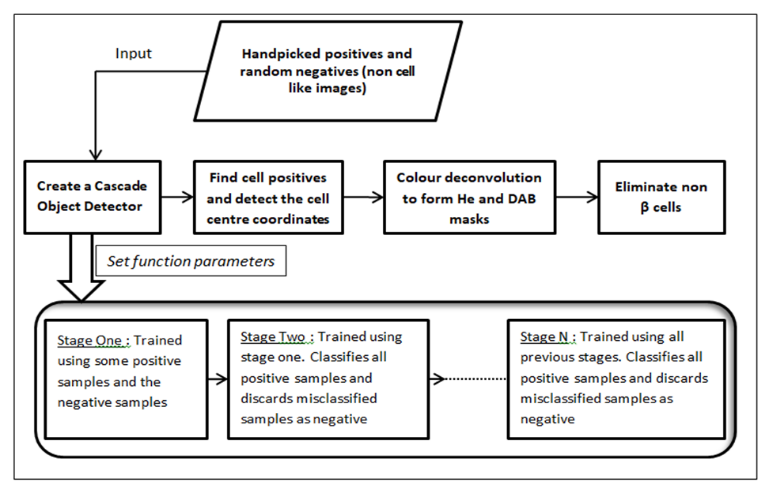
A decision tree typically starts with a single node, which branches into possible outcomes. Each of those outcomes leads to additional nodes, which branch off into other possibilities. It uses a graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. This gives it a treelike shape. Decision stumps are a special kind of decision trees having only one level, i.e, it consists of a single internal node connected to the terminal nodes. A training set of positive and negative images has to be prepared for the classification function. Along with the feature set this training set is used to learn a classification function.

Figure 1 : A schematic diagram of the proposed algorithm for β cell detection

For the learning of a decision stump, a feature and a value is selected which best splits the dataset into two parts by the selected feature according to some metrics. ID3 (Iterative Dichotomiser 3) algorithm[13]  is used to generate a decision tree from a dataset. It uses metrics like Entropy and Information gain to split the data.

Entropy *{\displaystyle H(S)}H(s)* is a measure of the amount of uncertainty in the (data) set *S* with *X* being the set of classes in *S* and *p(s)* being the proportion of number of elements in class x{\displaystyle x} to the number of elements in set S.

{\displaystyle S} {\displaystyle S}with wi

Entropy is calculated for each remaining attribute. The attribute with the smallest entropy is used to split the set *{\displaystyle S}S* on this iteration. The higher the entropy, the higher the potential to improve the classification.

Information gain *{\displaystyle IG(A)}IG(A)* is the measure of the difference in entropy from before to after the set *{\displaystyle S}S* is split on an attribute *{\displaystyle A}A*. The subsets created from splitting set *S* by attribute *A* are *T* such that

The proportion of the number of elements in {\displaystyle t}*t* to the number of elements in set {\displaystyle S}*S* is *p(t)* with *H(t)* being the entropy of subset *t*.

 Information gain can be calculated (instead of entropy) for each remaining attribute. The attribute with the largest information gain is used to split the set *S* on this iteration.

# Boosting the Detector

A variant of Adaptive Boosting (Adaboost) is used both to select a small set of features and to train the classifier. Freund *et al.*[14]proved that the training error of the strong classifier tends to zero exponentially in the number of rounds.

The algorithm can be shown as:

Initialize weight, = , , for yl = 0,1 respectively, where P+ and P- are the number of non ROIs and ROIs in the image set.

The algorithm is extended for an arbitrary number of rounds, l.

For i = 1,2,3,………..,l

1. Normalize the weights so that Wi,l is a probability distribution :

1. For each feature j, train a classifier hj which is restricted to using a single feature. The classifier’s error rate is evaluated with respect to .

=

1. Choose the classifier, , with lowest error . Update the weights:

The final classifier is:

The weak learning algorithm is designed to select the single rectangle feature which best separates the positive and negative examples. For each feature, the weak learner determines the optimal threshold classification function, such that the minimum number of examples are misclassified. The weak classifier can be mathematically expressed as the final classifier in the above algorithm.

Since decision trees are supervised learning algorithms, the need for a labelled training data becomes imperative. The labelled data is fed to the cascaded network and a classifier is developed which then takes test images(or a part thereof) as inputs and classifies them. Several hyperparameters can be tweaked to improve the performance.

# Cell Centre Detection

The classifier yields a number of Region of Interests(ROIs) as its output for a given test image input. The contrast of the cells is improved by employing histogram equalization. RGB ROI is converted to an inverted binary before computing the centroid of such a binary image. Morphological reconstruction is necessary for a satisfactory detection of centroid. Essentially a generalization of flood-filling, morphological reconstruction processes one image, called the marker, based on the characteristics of another image, called the mask. The high points, or peaks, in the marker image specify where processing begins. The peaks spread out, or dilate, while being forced to fit within the mask image. The spreading processing continues until the image values stop changing.

# Elimination of non βeta cells

At this stage we have all the cell centers detected. The next step in the algorithm is to eliminate the non β cells. A typical characteristic of β cells could be utilized for this. Since β cells have high insulin content, the stain by DAB-insulin is much more prominent in these cells. The algorithm used to achieve this is deconvolution of the colour information to compute the contribution of different stains as proposed by Ruifrok *et al.*[15] Here the optical density (OD) for R,G and B channels can be computed as

where *I0,C* is the intensity of light entering the specimen, *IC* is the intensity of light detected after passing the specimen, and subscript *C* indicates the detection channel. *A* is the amount of stain with absorption factor *C’*. In the case of three channels, the color system can be described in a matrix form which is then normalized to form a normalized *OD* matrix **M***.* If **C** is the 3 by 1 vector for amounts of the three stains at a particular pixel, then the vector of *OD* levels detected at that pixel is y=**CM**. From the above it is clear that **C**=M-1 [y]. This means, that multiplication of the OD image with the inverse of the OD matrix defined as the color-deconvolution matrix **D**, results in orthogonal representation of the stains forming the image; **C**=**D** [y].

**(d)**

Figure 2 : (a) Trained detector detecting all cell positives; (b) Centroid detection of all cell positives; (c) Original image and Masks for He and DAB stains; (d) Detected Beta cells

The corrected OD level values for DAB stain and Hematoxylin stains are formed by subtracting a portion of the red OD and the green OD from the enhanced blue OD to obtain the DAB OD and by subtracting a portion of the green OD and the blue OD from the enhanced red OD to obtain the hematoxylin OD respectively. The cells which have their DAB stain concentration below a certain threshold are detected as non β cells and hence eliminated.

# EXPERIMENTAL RESULTS

The proposed algorithm was implemented for the detection of β cells in Hematoxylin and DAB-insulin stained histology images. The dataset consisted of 20 images of mouse pancreatic sections stained with Hematoxylin and DAB-insulin and taken at 40×, the spatial resolution of each image being 1024×768 pixels. For a quantitative comparison of our algorithm we use four measures used in[16] and denote them as *μ*d*,* σd*, μ*n*,* σn. These measures evaluate the ability of the algorithm to identify centers and the total number of specialist(here, β) cells that were identified by the algorithm. *μ*d and σd denote the mean and standard deviation respectively of the closest distances for each of the identified cells, while *μ*n and σn denote the mean and standard deviation respectively of absolute difference of the number of nuclei between the ground truth and the result of automatic detection. The nature of specialist nuclei detection in the presence of an often subjective *ground truth* is challenging. The ground truth was obtained by fusing the markings provided by three experts. For each of the images in the dataset, three such images (one image per expert) containing round blobs around the marked cell centers are combined by adding them together and then thresholding, resulting in a fused ground truth (FGT).

For comparison purposes, we tested the performance of the proposed Ensemble of Boosted Decision Stumps(EoBDS ) against Local isotropic phase symmetry measure(LIPSyM) as proposed by Kuse *et al.*[17]and also against the standard Laplacian of Gaussian(LoG) filter as proposed by Byun *et al.*[18]. The results (Table 1) depicts that the proposed method outperforms LIPSyM as well as LoG filter.

Another significant aspect for the utility of such an algorithm is the time taken for the implementation on a single image. This lends an idea about its applicability in a real life scenario. As discussed above the human marking procedure is a long and tedious one and depends on several human factors such as fatigue level of the human, expertise of the human etc. As an example, an experiment consisting of 6000 images of the size 1024×768, each containing 50-300 beta cells takes about 500 man hours at the rate of 4-5 minutes per image. LIPSyM algorithm takes about 50 computer hours without any manual intervention at the rate of 30 seconds per image. LIPSyM algorithm was implemented in Matlab and it worked on a 2.66GHz Quad-Core Workstation. The proposed algorithm, EoBDS was also implemented in Matlab. Each of the 30 staged cascade detectors took 40 minutes to train. It took 50 computer hours at the rate of about 30 seconds per image to mark all the beta cells. We worked on a 1.70GHz Dual-Core machine unlike the LIPSyM algorithm implementer. Quite evidently, the proposed algorithm is faster and more accurate and can be used in real life circumstances.

# DISCUSSIONS

The dataset contains twenty 1024×768 pixel images. 20 fused ground truth images contain the β cell coordinates. However, one of the critical parameters for the creation of a satisfactory cascade detector is the number of cascaded stages trained. Increasing the number of stages may result in a more accurate detector, however, it also increases training time. More stages require more training images, because at each stage, some number of positive and negative samples are eliminated. The false alarm rate is the fraction of negative training samples incorrectly classified as positive samples. The true positive rate is the fraction of correctly classified positive training samples. The number of stages depends on the values of false alarm rate and true positive rate which too are hyperparameters. More stages can also lead to increase the false alarm rate. Higher values of false alarm rate reduce the complexity in each stage.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **μd** | **σd** | **μn** | **σn** |
| LoG | 9.21 | 2.83 | 15.65 | 20.30 |
| LIPSyM | 7.82 | 2.18 | 10.35 | 9.39 |
| EoBDS | **7.11** | **1.38** | **7.40** | **6.67** |

As is obvious from the above discussion, the number of training images have to be significant enough to have a reasonable detection accuracy. The dataset available does not provide us with enough training data to produce a single good enough detector. To overcome this we first manually pick out the ROIs for all cells (β as well as non β) in an image using an Image Labeller. Now we train and build 4 such detectors. For the 1st detector, we choose the first 5 images as the test dataset and use the last 15 to train the detector. Similarly, we train the other three detectors using non intersecting sets of 5 test images and 15 training images from the dataset. Finally all the four measures computed from the 20 images are averaged and put up for comparison.

Table 1 : Experimental results for β cell detection using different algorithms (Best results in bold)

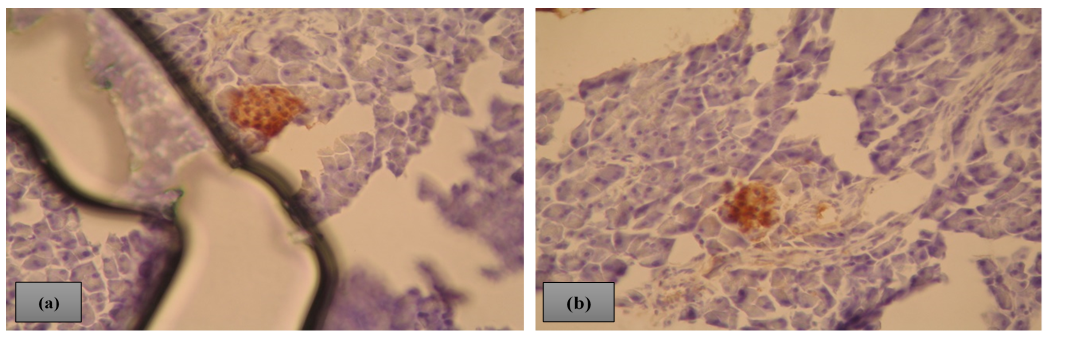
One more hurdle we face while eliminating the non β cells is that the concentration of DAB stain is found to be significant along the boundary edges. A pictorial representation is shown in Fig 2(c) and mentioned region is enclosed in an ellipse. To overcome this situation we make use of the visual interpretation that such edges have a greater concentration of Hematoxylin stain than DAB stain and that all other valid DAB stained regions having β cells don’t show a similar pattern. Hence, by comparing the Hematoxylin stain to the DAB stain we could remove the ambiguity along the boundaries.

Figure 3 : Images with aberrant characteristics troubling the detector

# CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a novel measure of employing an ensemble of boosted decision stumps to create a detector for cell identification. We employ EoBDS in a systematic framework for labelling particular stain-specific cells such as insulin-rich β cells in pancreatic sections. Counting β cell mass plays a critical role in studying the development and cure of diabetes in mouse models. The results of the proposed method depicted its effectiveness.

However, these results can be improved to a further extent. The images in Fig.3 account for such high value for μd and σd as the trained detector never encountered such anomaly before.

In Fig 3(a) there is a huge intensity difference at the centre and since the detector is based on Haar features, many cells are falsely identified along the dark boundary. In Fig 3(b), at the far upper left, two faint stained cells are present. However, these are missed by the detector and hence the mean distance increases by a vigorous amount. If the detector is trained on a much larger(say, around 100 images) and a much varied image set possessing several anomalies, it can be fairly predicted that the results would improve. Also, different machine learning algorithms can be implemented, with or without tweaking their hyperparameters, to obtain a better performance.

References

1. Butler AE, Janson J, Soeller WC, Butler PC. Increased β-cell apoptosis prevents adaptive increase in β-cell mass in mouse model of type 2 diabetes. Diabetes 2003;52:2304-14.
2. Xiong G, Zhou X, Ji L. Automated segmentation of drosophila RNAi fluorescence cellular images using deformable models. Circuits and Systems I: Regular Papers, IEEE Transactions on, 2006;53:2415-24, 2006.
3. Hafiane A, Bunyak F, Palaniappan K. Clustering initiated multiphase active contours and robust separation of nuclei groups for tissue segmentation. In Pattern Recognition, 2008. ICPR 2008. 19th International Conference on, IEEE, 2008. p. 1-4.
4. Hafiane A, Bunyak F, Palaniappan K. Fuzzy clustering and active contours for histopathology image segmentation and nuclei detection. In Advanced Concepts for Intelligent Vision Systems. Berlin, Heidelberg: Springer; 2008. p. 903-14.
5. Bamford P, Lovell B. Unsupervised cell nucleus segmentation with active contours. Signal Process 1998;71:203-13.
6. Hu M, Ping X, Ding Y. A new active contour model and its application on cell segmentation. In Control, Automation, Robotics and Vision Conference, 2004. ICARCV 2004 8th. IEEE 2004;2:1104-7.
7. MacEachern LA, Manku T. Genetic algorithms for active contour optimization. In Circuits and Systems, 1998. ISCAS’98. Proceedings of the 1998 IEEE International Symposium on, IEEE 1998;4:229-32.
8. Mouroutis T, Roberts SJ, Bharath AA. Robust cell nuclei segmentation using statistical modelling. Bioimaging 1998;6:79-91.
9. Gurcan MN, Pan T, Shimada H, Saltz J. Image analysis for neuroblastoma classification: Segmentation of cell nuclei. In Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE, pages 4844-4847. IEEE, 2006.
10. Pelengaris S, Khan M, Evan GI. Suppression of myc-induced apoptosis in β cells exposes multiple oncogenic properties of myc and triggers carcinogenic progression. Cell 202;109:321-34.
11. Viola, P., and M. J. Jones. "Rapid Object Detection using a Boosted Cascade of Simple Features." Proceedings of the 2001 IEEE Computer Society Conference. Volume 1, 15 April 2001, pp. I-511–I-518.
12. C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *International Conference on* *Computer Vision*, 1998.
13. Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106
14. Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *Computational Learning Theory: Eurocolt ’95*, pages 23–37. Springer-Verlag, 1995.
15. Ruifrok AC, Johnston DA. Quantification of histological staining by color deconvolution. Anal Quant Cytol Histol 23: 291-299, 2001.
16. Gurcan M, Madabhushi A, Rajpoot N. Pattern recognition in histopathological images: An ICPR 2010 contest. Recognizing Patterns in Signals, Speech, Images and Videos, pages 226-234, 2010.
17. Kuse M, Wang Y, Kalasannavar V, Khan M, Rajpoot N. Local isotropic phase symmetry measure for detection of beta cells and lymphocytes. J Pathol Inform 2011;2:2.
18. Byun J, Verardo MR, Sumengen B, Lewis GP, Manjunath BS, Fisher SK. Automated tool for the detection of cell nuclei in digital microscopic images: Application to retinal images. Mol Vis 2006;12:949-60.