A Machine Learning Approach Towards Fatty Liver Disease Detection in Liver Ultrasound Images

Adarsh Kuzhipathalil[1], Anto Thomas[1], Keerthana Chand[1], Elmer Jeto Gomes Ataide[2], Alexander Link[4], Annika Niemann[3,5], Sylvia Saalfeld[3,5], Michael Friebe[2], Jens Ziegle[2]

[1]Faculty of Computer Science, University of Magdeburg, Germany
[2]INKA-Application Driven Research, University of Magdeburg, Germany
[3]Department of Simulation and Graphics, University of Magdeburg, Germany
[4]Department of Gastroenterology, Hepatology and Infectious Diseases, University of Magdeburg
[5]STIMULATE Research Campus, University of Magdeburg
adarsh.kuzhipathalil@st.ovgu.de

Abstract. Fatty liver disease (FLD) is one of the prominent diseases which affects the normal functionality of the liver by building vacuoles of fat in the liver cells. FLD is an indicator of imbalance in the metabolic system and could cause cardiovascular diseases, liver inflammation, cirrhosis and furthermore neoplasm. Detection and specification of a FLD are beneficial to arrange an early and best adapted treatment. We present a computer aided diagnostic (CAD) tool for FLD detection using ultrasound images. The developed pipeline consists of separate segmentation and classification modules. During the development phase these modules were trained on 6 patient cases and validated with 2. The whole model was evaluated on a totally different set of data with 5 patient cases and performed with an overall classification accuracy of 0.84. The model showed impressive performance considering the size of training data. Also the multi-module architecture enables the predictions from the model to be better explainable.

1 Introduction

The liver has a significant role in detoxification protein synthesis and production of chemicals that helps in proper digestion. There are few diseases which affects its notmal functionality. Examples of liver disease include: Fatty liver disease (FLD), cirrhosis and cancer. Our study mainly concentrates on FLD. Blood tests, liver ultrasound (US) imaging, computed tomography and liver biopsy can be employed to detect and diagnose FLD [1].

US imaging is more common due to its lower cost and its non-invasive nature. FLD diagnosis often depends of the subjective judgement of the physicans due to differences in US equipment, poor image quality and the physical differences of patients. As physicians depend on different aspects of the US image for

diagnosis, there are high variations in diagnosis which often depends on the experience of the physicians. Incorrect diagnosis could lead to wrong and contra-productive treatment, e.g. wrong concentration of medication may harm the patient. Previously published works on FLD [2] and diffused liver disease [3] detection in US images was successful in developing computer aided diagnostic (CAD) tools which aided the physicians in the diagnosis. Both of these works gives out one single prediction for the whole US image. We developed the tool which help in detecting fatty tissue in the US image and generate the prediction in such a way that it is better explainable for physicians with different levels of experience.

2 Material and methods

The major distinguishing character is the difference in texture patterns between fatty and non-fatty liver observed from US images. The FLD detection problem can essentially be modelled as a pattern recognition problem by extracting the texture features from respective classes and further building the classifier. There are previous efforts in developing similar pipelines for liver US images. Fractal Dimension Texture Analysis (FDTA), the Spatial Gray Level Dependence Matrices (SGLDM), the Gray Level Difference Statistics (GLDS), the Gray Level Run Length Statistics (RUNL), First Order Gray Level Parameters (FOP), Gray level Co-occurrence matrix (GLCM) and Wavelet Transforms (WT) are some commonly used texture feature extraction techniques [3, 4].

For developing the FLD detection pipeline, we have generated a dataset consisting of 13 patient cases in total. In the course of development 8 out of 13 were used and the remaining 5 were used to test the system at the fully developed stage. The study was performed according to the "World Medical Association Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects." All subjects included in the study provided written informed consent and the study protocol was approved by the local Institutional Review Board of Otto von Guericke University, Magdeburg. All images were anonymized prior to analysis.

The dataset was generated by two US systems (Philips iU22 and GE LOGIQ E10) consisting 964 images in total in fatty class and 1060 images in non-fatty class. The images were stored in JPEG format. The datasets were labelled according to the clinical diagnostic outcome. Segmentation ground truths were generated with the help of an experienced hepatologist. These were chosen in such a way that these contained only the characteristic liver texture excluding vessels and border tissues. This means that the segmentation masks did not necessarily contain the whole liver parts in the US image. Also the non fatty regions inside the fatty liver samples were not considered differently while generating the segmentation masks. This was done to enable a pipeline to classify similar regions without explicitly training on such regions. Furthermore, this dataset was used in the machine learning (ML) pipeline for feature extraction and classification.

2.1 Training and output pipelines

The training phase of the pipeline is split in two different branches (Fig 1). The input dataset consists of fatty and non-fatty liver US images and the corresponding segmentation masks. The first branch takes the data and converts it to liver and background classes. The liver and background classes along with the labels makes up the training data for the first module. One the other branch, just the region of interest is extracted using the segmentation masks. Which means only the liver portion of the US image is selected and grouped as fatty and non-fatty class. These two classes along with the corresponding labels are used as the training data for the second module.

Before the feature extraction was performed on the US images, the images were converted into patches. The texture features were extracted from these generated patches since a global feature vector from the whole image is inefficient in case of US image texture, as the texture has large variations throughout the image. Overlapping patches were used in this case. This helped in better usage of the available data. The optimal patch size was found out to be 30×30 pixels. This was evaluated by experimenting with multiple patch sizes ranging from 10×10 till 50×50 pixels. Smaller patches failed to extract the texture characteristics and on the other hand larger patches resulted in extracting features from a mix of different textures.

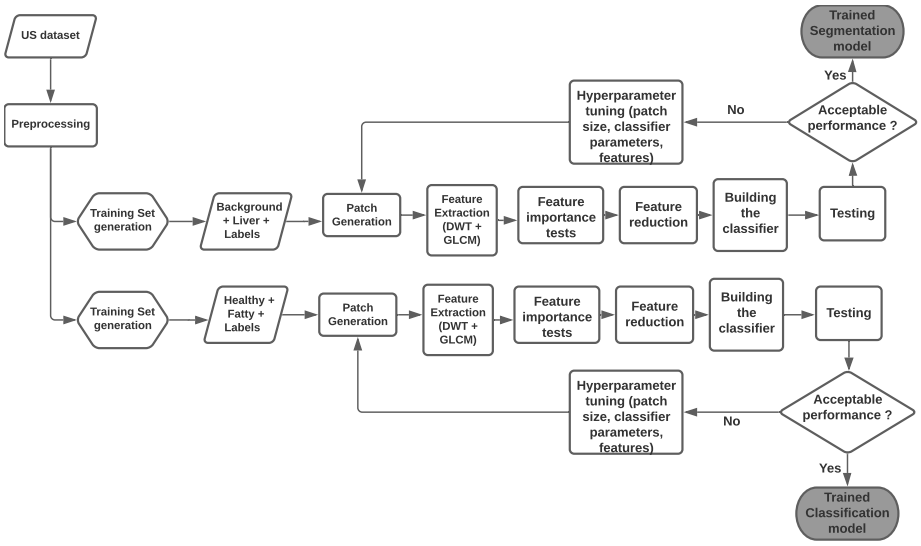
For the prediction and visualization (Fig. 2), the input image is first given to the segmentation module and then to classification module to produce the final prediction results. For both the pipelines, feature importance studies and feature reduction methods using Principal Component Analysis (PCA) were performed. For the segmentation and classification modules random forest classifier was trained with the corresponding set of reduced features.

2.2 Texture feature extraction

Feature extraction step in the pattern recognition problem is the crucial and most influential step of the whole pipeline as it encodes the input image. The encoded features were used in the further decision process. The choice of features directly impacts the accuracy and the generalization of the classes of interest, in this problem the non-fatty and fatty liver tissues. In our approach we used 90 features in total. The majority of features (84) were extracted from Discrete Wavelet Transform (DWT) methods and 6 features from the GLCM.

The discrete wavelet transform (DWT) is performed using a set of 7 discrete wavelets namely Daubechies, Haar, Biorthogonal, Reversebiorthogonal, Symlets, Coiflets, and Mayer wavelets. These wavelets have different scales and translations. DWT decomposes the signal into mutually orthogonal sets of wavelets. The mean, variance and entropy were calculated for each patch decomposed using the seven wavelets, which makes up 84 features of feature vector. The contrast, dissimilarity, homogeneity, Angular Second Moment, energy, and correlation features (6 features in total) were extracted from the GLCM.

Fig. 1. Training pipeline.



2.3 Model evaluation

Two out of eight patient cases were chosen as validation dataset to evaluate the performance of the model. The modules were evaluated separately using this data. Finally, to check the qualitative performance of the model pipeline and the combined performance of individual modules, 5 patient test cases were given to the pipeline and evaluated. This was done as those 5 cases contained a mix of different FLD stages. The qualitative visual inspection of the predictions mainly concentrated on the accuracy and completeness of segmentation, segmentation performance around the vessels and the classification performance. To evaluate the whole model on the binary classification problem (fatty or non-fatty), a prediction ratio (PR) was generated. PR is the ratio of patches which were classified as fatty to the ones which were classified as non-fatty. This was used as a quantitative measure to test the performance of the whole system.

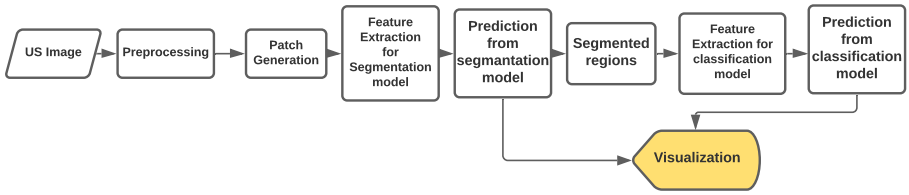


Fig. 2. Prediction pipeline.

Table 1. Performance matrix of the segmentation model.

	Precision	Recall	f1 Score	Support
Background	0.81	0.78	0.80	81,200
Liver	0.79	0.82	0.80	81,200
Accuracy			0.81	162,400

Table 2. Performance matrix of the classification model.

	Precision	Recall	f1 Score	Support
Non-Fatty	0.70	0.72	0.71	36,000
Fatty	0.74	0.74	0.70	45,200
Accuracy			0.71	81,200

3 Results

The segmentation module was tested with 162,400 patches and was able to achieve an accuracy of 0.81 (Table 1). Then the classification module was tested with the patches extracted from the liver sections. The module was tested with 81,200 patches and was able to achieve an average accuracy of 0.71 (Table 2). Equal number of patches were sampled from each of the classes while training and testing. This was necessary since the background is greater in size when compared to the liver sections.

The model was tested with a set of unseen samples for evaluating the qualitative performances. Fig 4 shows the predictions of a fatty sample when tested on the pipeline. The model was able to produce well generalized and accurate results. As seen in the figure, the ground truth contains only a portion of the liver in the US image, but the model was able to segment and classify most of the portions of the liver beyond the ground truth. Similar behavior is shown by the model on all of the testing images. The model performances with different thresholds for PR were evaluated. The model's binary classification accuracy was 0.84 with a PR of 0.60. The model predictions are helpful in locating the fatty regions for further diagnosis.

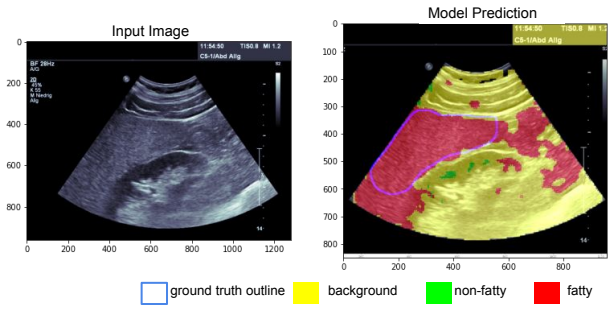
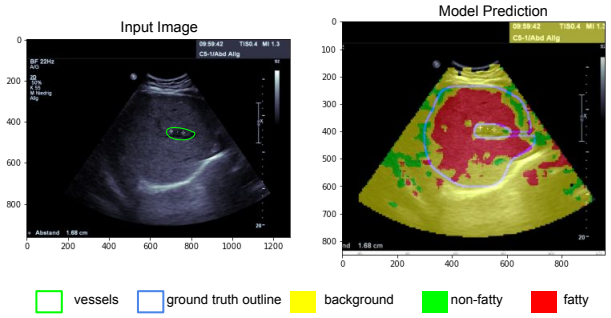


Fig. 3. Model prediction for an input image with a fatty sample.

Fig. 4. Segmentation around the vessels.



4 Discussions

The fatty tissue classification worked well using a pipeline consisting of segmentation and classification modules. Our proposed multi stage classification solution with the patch based approach was successful in the tissue texture recognition for our datasets. As depicted in Figure 4, the model was capable of successfully processing images that include vessels and learned to segment liver texture around vessels. Due to the adopted strategy in selection of ground truth, the quantitative model evaluation doesn't consider the regions which were predicted outside the ground truth. To evaluating model performance on these regions, a different set of ground truth with the segmentation masks covering the whole liver portion has to be generated. Due to the adopted strategy in generating the segmentation masks, model is not explicitly trained on minor variations in the texture. Hence the chances of overfitting is minimized.

There are few drawbacks for the developed model. Both the segmentation and classification modules were sensitive to the image scale. The used dataset was a mix of different scaled images. But rescaling the images to a single size resulted in loosing the details as our datasets were in JPEG format. Using DICOM or TIFF files would give better results.

References

1. Langer SG, Carter SJ, Haynor DR, et al. Image acquisition: ultrasound, computed tomography, and magnetic resonance imaging. World J Surg. 2001;25:1428–1437.
2. Fu M, D S, Hussain. Automated classification of liver disorders using ultrasound images. J Med Syst. 2012;36(5):3163–3172.
3. Kyriacou E, Pavlopoulos S, et al. Computer assisted characterization of diffused liver disease using image texture analysis techniques on B-scan images. IEEE Nucl Sci Symp Conf Rec (1997). 1997;2:1479–1483.
4. Kyriacou E, Pavlopoulos S, et al. Fuzzy neural network-based texture analysis of ultrasonic images. IEEE Eng Med Biol Mag. 2000;19:39–47.