UNIVERSITY OF Southampton

Machine Learning for Liver Disease Classification

Master of Philosophy in Computer Science

Benjamin D. Jesty

Saturday 25th May, 2019

Supervisors: Prof. M. Niranjan, Prof. S. Dasmahapatra

Department of Electronics and Computer Science, University of Southampton

Abstract

In this work the use of machine learning in medicine, with a particular focus on liver disease, is investigated and summarised. A variety of machine learning techniques for feature selection and classification are then applied to a novel medical application. A dataset of healthy (20,089) and unhealthy (714) patients' full blood count blood tests is analysed to further medical understanding of how liver disease affects the blood and to enable a new diagnosis technique based on commonly available information. Methods for outlier identification and robust classification are also introduced and evaluated.

Logistic regression and soft margin support vector machines are used to classify patients as healthy or unhealthy based on the blood tests. Feature selection is performed on the data. Three primary features (90% area under receiver operating characteristic curve accuracy) and four secondary features are found for the peak accuracy based on the 7-feature support vector machine classifier of $92 \pm 0.5\%$. These features are verified by a liver specialist to be influenced by liver disease. The final classifier is further tested on a completely new dataset of 100,000 patients' data and achieved 90% accuracy, marginally outperforming the classifier designed by a liver specialist.

Feature selection and classification tasks are performed on time cohorts to investigate temporal information in the data. Differences in features selected are found between blood tests taken near diagnosis and years prior. Classification accuracy is shown to decrease steadily as time prior to diagnosis increases. However, blood tests taken 6 years prior to diagnosis can still be dichotomised with greater than 75% accuracy.

An outlier rejecting support vector machine is developed and tested on artificial datasets and the portal hypertension dataset. The outlier rejection during training shows major improvements for small, well structured datasets but struggles to improve on soft margin support vector machines for larger, more complex datasets.

Contents

Research Thesis: Declaration of Authorship

Name: Benjamin D. Jesty
Title of thesis: Machine Learning for Liver Disease Classification

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signature:
Date:

Acknowledgements

I would like to thank Prof. Nick Sheron for his work in collecting, anonymising and sharing data on liver disease patients, in particular the portal hypertension dataset which formed the basis for much of my research. His guidance on the medical aspects of the research has also been invaluable.

I would also like to thank my supervisors, Prof. Mahesan Niranjan and Prof. Srinandan Dasmahapatra, for their support both academically and personally over the course of this research project.

Finally, I would like to thank my family, my church and my Lord Jesus Christ for keeping me going through some of the most challenging times of my life so far.

Chapter 1

Machine Learning in Medicine

The use of machine learning as a tool to aid medical diagnostics is becoming increasingly prevalent. The exact scenarios in which machine learning methodologies are applied are very varied. However, they all have one thing in common: rapidly escalating access to large quantities of data. This chapter gives a brief overview of some of the primary challenges and existing solutions to machine learning problems in medicine.

1.1 Feature Selection

Feature selection is a method to reduce model complexity by selecting a subset of the most informative features in the input space and rejecting the remaining features. Reducing the number of features used in a model has three main benefits:

1. The model is easier to interpret
2. The model is trained and tested on less data so processing time is reduced
3. The model is less likely to overfit a training dataset

1.1.1 Preventing Overfitting

A model is considered to have overfitted a set of data if it performs significantly better on the data it was trained on (training set) than a separate set of previously unseen data of the same origin (testing set). This happens

when a system learns features specific to the training set (often statistically random correlations) which are not found in the general data population. Models are most susceptible to overfitting when they have many features, P, and few examples, N. It is generally best practice to have enough data points that $N >> P$. Situations where this is not possible (e.g. genomics, computer vision) have a greater focus on narrowing down the input features, either by exclusion or combination into higher-level features.

1.1.2 Independence of Features

When selecting the best feature subset one must consider the independence of the individual features. It is rarely the case that the best set of n features is simply the combination of the best n individual features. If any subset[1] of features, A, can predict the values of another subset of features, B, the combination of the two subsets, {A,B}, will be no more valuable than A alone.

For example, imagine a case where the aim is to estimate the value of a house based on F = total floor area, R = number of rooms and C = condition of the house. F and R both give information about the size of the house, so individually they will be good predictors. C alone will be a comparatively poor predictor since size is more important than condition for house prices. If one were to combine the top two predictors however, the resulting predictor {F,R} would be little better than F or R alone, since either can estimate the other with reasonable confidence. Combining F or R with C is much more likely to create a better classifier than {F,R} since C is independent of both F and R.

Due to this, the feature subsets found by any feature selection algorithm other than a brute force search can never be assumed to contain the best individual predictors or even the best smaller subsets of predictors.

1.1.3 Feature Selection Method Summary

There are many methods used for feature selection. The ideal scenario would be to test a model built on every possible distinct combination of features. However, this is infeasible for all but the smallest datasets since

[1] a subset could be a single feature

the time complexity of this exhaustive search is $O((n+1)!)$ where n is the number of features present. The choice of which method to use in a given scenario depends on the requirements of accuracy and efficiency. Some algorithms (such as greedy forward feature selection, GFFS, described in section 3.3.3) prioritise accuracy over efficiency by comparing many selective feature combinations, though still much fewer than exhaustive search. The time complexity for GFFS is $O((n+1)^2)$. These techniques tend to be limited to work on datasets with a low number of features. Large datasets, such as work on bioinformatics and computer vision, can easily have in excess of a million features and it is therefore infeasible running $O(n^2)$ algorithms on them. At the other end of the scale, heuristic techniques aim to find good estimates of the ideal feature subset in minimal time. At the extreme efficiency end of the spectrum there are techniques that predict the ranking of features based on a single test, such as ranking by the magnitude of the feature weights given by a linear SVM, which have time complexity of $O(1)$. Though heuristic techniques often do not find subsets of features as close to the theoretical optimum, they are still widely used in large datasets where time complexity is an issue (Saeys, Inza, and Larrañaga, 2007).

A few examples of feature selection techniques (from slowest to fastest) are given below:

- Brute force - Complexity $O((n+1)!)$
  Every possible subset of features is evaluated and the best subset is retained.
- Sequential search - Complexity $O((n+1)^2)$
  Forward selection: feature set S starts off empty, set of unused features U starts off containing all features. On each iteration: evaluate the performance of $f(S \cup x)$ for each $x \in U$ and move the best performing feature x from U to S. Repeat until a high enough accuracy is reached or the required number of features are selected.
  Backward elimination: feature set S starts off with all features. On each iteration: evaluate the performance of $f(S \setminus x)$ for each $x \in S$ and remove the feature x that reduced the performance least from S. Repeat until a minimum accuracy threshold is reached or the required number of features are removed.
- Sequential ranked search - Complexity $O(n)$

Each feature is ranked (see below), then features are added one by one based on their rank until the required accuracy is reached.

- Ranking methods - Complexity $O(1)$
  Each feature is given a ranking (such as Pearson's Correlation Coefficient) and the top n features are taken based on their rank, where n is the required number of features.

1.2 Accuracy vs Interpretability

In machine learning there is a general trade-off between the accuracy and interpretability of a system (Caruana et al., 2015). On the most interpretable end of the scale, linear systems give a single scalar weight to each input feature. It is therefore easy to see how the system reached its conclusion - the higher the magnitude of each feature's weight, the more influence it has on the final conclusion. The cost of such a system is that linear systems are often not capable of picking up on more complex underlying patterns in the data and therefore suffer in terms of accuracy. On the other end of the scale are systems like deep recurrent neural networks. Such systems often achieve the highest accuracies possible with current understanding since the internal state of the abstract networks can pick up on more complex underlying patterns in data. However, due to the complexity of the networks generated it is often unclear exactly how and why the system gives the answer it gives. (Choi et al., 2016)

Both interpretability and accuracy are important in medical applications; accurate systems are crucial for correct diagnoses and treatments to be given and interpretable systems allow researchers to develop their understanding of the medical processes they tackle, as well as allowing doctors to correct false assumptions in the system. The machine learning techniques chosen for a project will be heavily influenced by the balance required between accuracy and interpretability.

1.3 Types of Machine Learning used

Medical machine learning tasks can be roughly grouped into five categories: genomics, audio analysis, computer vision, natural language processing and

direct health record regression. These groups are by no means exclusive (tasks may draw from multiple groups) or exhaustive (tasks may have novel requirements outside of these groups) but they should serve to highlight the main challenges faced by different tasks and the resulting solutions.

1.3.1 Genomics

Genomics is the study of the structure and function of the entire set of DNA (the genome) of an organism. Sections of the genome, known as genes, define how proteins are produced (through multiple intermediate stages). These proteins in turn define both the physical structure of the organism and also its chemical signalling (and consequently certain behaviours). Identifying genes within the genome, along with mapping genotypes (the physical structure of the gene's section of DNA) to phenotypes (observable characteristics such as physical structure and behaviour) has great medical value. For example, in the study of the human genome finding genotypes that cause medical disorders could allow early diagnosis and development of preventative medication for the disorders, sometimes before external symptoms are even displayed. Genomics can also be used to sequence the genomes of organisms behind infectious diseases, allowing scientists a better understanding of how to best treat or even eradicate the disease.

Genomes encode a very large amount of data; the human genome, for example, has approximately 3 billion base pairs (features), each of which can take one of four values. Analysis of such large quantities of data lends itself nicely to techniques developed in machine learning, such as those reviewed by Leung et al., 2016.

Since there are so many features (base-pairs) in genomes and often very few instances (genomes) to compare, machine learning techniques have to be designed with measures to prevent overfitting (see 1.1.1) when dealing with genomics data. One crucial step in this process is feature selection, since a model with fewer features will be less prone to overfitting than a model with a greater number of features. Another technique to establish whether a model is overfitting the data is to train the model on randomised data, with the expectation that the model should find a pattern in the real data but find nothing in the randomised version. If the model claims to find similar strength trends in random data compared to those found in the real data,

the model is most likely overfitting.

1.3.2 Computer Vision

Computer vision is the use of computer algorithms to identify useful information in digital images and videos. Some applications aim to re-create human visual abilities (e.g. face detection). Others go beyond the capability of human vision, such as using images composed from wavelengths of the electromagnetic spectrum humans cannot see (e.g. radar and magnetic resonance imaging (MRI) machines). Many applications use a combination of the two (e.g. self-driving cars which use visible light in combination with radar, lidar, etc).

Machine learning is used extensively in computer vision for classification and clustering (including labelling). The primary challenge in using machine learning in this field is the sheer quantity of information. Each pixel in an input image represents a single feature. Even a small image of dimensions 100x100 pixels represents 10,000 input features (30,000 for a colour image), each of which typically have 256 possible values. It is therefore infeasible to have more samples than features ($N >> P$), which means a lot of work goes into extracting higher-level features. These features are typically values found by kernel functions (masks) applied to groups of nearby pixels. The high-level features can then be used directly for analysis or they can be passed into higher-level feature extractors, as is the case in convolutional neural networks, before eventual analysis.

Applied in medicine, there is a distinction to be made between image augmentation to assist human analysis and direct image analysis that attempts to automate an entire process independent of human intervention. Two examples to illustrate this distinction are the works of Mahmud et al., 2015 and Havaei et al., 2017 respectively.

In the work of Mahmud et al., 2015 the authors describe a system to assist doctors performing gastrointestinal (GI) endoscopies. In GI endoscopies a high-definition camera is inserted on a flexible endoscope into a patient's intestines. The endoscopist uses a live feed from the camera to guide the endoscope through the intestines while simultaneously attempting to identify defects (such as polyps) on the intestinal walls that may be causing medical issues. The authors highlight the issue that endoscopists frequently

miss defects even when they show up on the camera. They suggest a new system whereby the live video is analysed by a computer to identify likely polyp sites and highlight those areas in real-time on the endoscopist's screen. (Mahmud et al., 2015)

This example shows how computer vision can be designed into systems to enhance the work of specialists. This methodology has the advantage of not needing to be as accurate as direct analysis techniques, since it only offers suggestions that the specialist will rule on, rather than aiming giving final diagnoses.

In comparison, Havaei et al., 2017 introduce an example of a direct-analysis system. In their work, the authors develop a deep neural network to segment cross-sectional images of brains taken from MRI machines. The network is trained to pick up on glioblastomas (a type of brain cancer). The network can then automatically classify regions of the brain scan images without necessary intervention from medical professionals. (Havaei et al., 2017)

This methodology has the benefit of requiring minimal specialist knowledge to use the developed system (see also Onu et al., 2017) but the consequent disadvantage of needing to be much more thoroughly tested since it attempts to give a final diagnosis rather than just a suggestion.

1.3.3 Audio Analysis

Audio analysis seeks to find interpretable patterns in digital audio recordings. Typical applications include speech recognition, music identification and detecting known marker sounds against a background noise, for example to detect the presence of a certain animal in a forest. Auditory information can be viewed as a spectrogram (see Figure 1.1) and therefore audio analysis becomes a special case of computer vision, drawing on the extensive suite of techniques already present in that field.

An example use of machine learning-based audio analysis in medicine is given in the work of Onu et al., 2017. The authors set out to tackle the issue of infant mortality caused by birth asphyxia in "resource-poor settings" which do not have access to expensive or highly-technical diagnostic methods. They use Support Vector Machines (SVMs) combined with existing

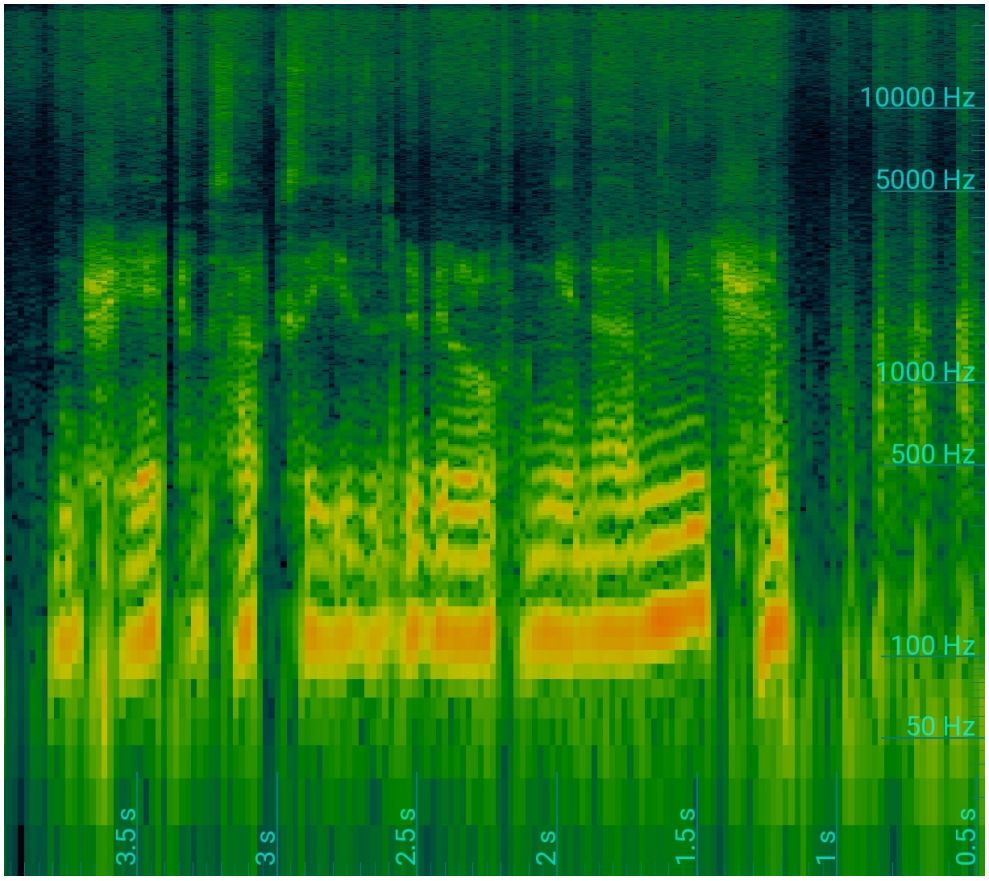


Figure 1.1: Spectrogram of author saying "Machine Learning for Liver Disease Classification". Brightness indicates a level of a certain frequency (y-axis) at a given time (x-axis). Scale: orange (highest) - yellow - green - black (lowest).

speech recognition techniques to design a classifier that can distinguish the cry of a healthy baby from that of a baby with birth asphyxia. This classifier can then be used on a mobile phone with its built-in microphone to give immediate, low-cost and low-expertise diagnoses.

1.3.4 Natural Language Processing

Natural language processing attempts to extract useful information from text written with a particular set of structural rules (a language). Machine learning techniques are used to give numerical values to measurable features, such as word or phrase frequencies. These values can then be analysed using regression tools.

A passage of text will exhibit patterns based not just on the national language of the author but also on the purpose of the text. For example, a transcribed informal conversation between close friends will have a different style to an academic paper, even if both are written in English. Any analysis on text therefore needs to be normalised against the background specific to that form of text.

Within patient health records is a wealth of written reports. These often follow common structures (statements of diagnoses, clinical narratives) and contain large quantities of diagnostically useful information. Developing NLP systems capable of unlocking the data stored in this text would therefore provide an influx of machine learning-processable data from historical records that could improve future diagnostic systems and care. (Pons et al., 2016)

1.3.5 Health Record Regression

Regression analysis seeks to establish what underlying relationships exist between variables (features). The aim is to find a function, f, such that $f(x) = y$ where x is the vector of known feature values (predictors) and y is the dependent feature value. For traditional regression systems both x and y are continuous-valued features, however modifications can be made to allow discrete inputs. Classification systems simply use one or more threshold values on the predicted value of y. These threshold values can either be learned as part of the training process or predefined, depending on the classification technique employed.

Regression analysis will be used at some stage in all of the prior example groups. However, each of those cases require extra processing (particularly feature selection) beforehand. Numerical datasets compiled from patients' health records can commonly be analysed with significantly less, if any, preprocessing.

An example use of regression in healthcare is introduced in the work of Shameer et al., 2017. In this paper, the authors used the Naive Bayes algorithm to find the relationships between 4,205 input features taken from patients' health records and the patients' readmittance rate. This system would then allow doctors to identify which patients need extra attention prior to being discharged.

An example of classification from health record data is given by Patel and Joshi, 2013. This work looks at classifying heart disease directly from existing health record information. This system can then identify potential heart disease in patients from past data without them needing to be sent for specific tests for heart disease, improving early detection.

Chapter 2

Machine Learning for Liver Disease Analysis

In this chapter the past use of machine learning in the classification and diagnosis of liver disease is reviewed. Work in this field to date can be grouped into three areas:

- Image analysis of MRI, CT or ultrasound scans
- Mortality prediction among hepatitis patients
- Diagnosis of liver diseases from numeric and binary data

Many of the reviewed studies here used one or more datasets freely available from the Machine Learning Repository of the University of California, Irvine[1]. These will be referred to as UCI ML datasets.

2.1 Image Analysis

The work reviewed in this section all focusses on the analysis of medical images. Though the sources of images differ between studies, the methods used to analyse them are comparable.

2.1.1 Data

Guo et al., 2009 compile a dataset of magnetic resonance images (MRIs) of 40 rat livers, some with hepatocellular carcinomas (HCCs). 106 images (82

[1]https://archive.ics.uci.edu/ml/index.php (page last accessed 30/10/2018)

HCC regions, 24 non-HCC) are split into 161 regions of interest (ROI) (81 HCC, 80 non-HCC).

Mala, Sadasivam, and Alagappan, 2015 use a dataset of computerised tomography (CT) scans of human livers. 40 CT scans were performed on patients with fatty liver disease and 40 on patients with cirrhotic patients and each scan has approximately 20 slices (individual images).

Virmani et al., 2013 use a dataset of 56 ultrasound images; 15 of healthy livers, 16 of cirrhotic livers and 25 of livers with HCC. Each ultrasound was taken from a separate patient. 180 non-overlapping ROIs are extracted by a radiologist, 60 per class, as the final dataset.

2.1.2 Methods

All three studies used wavelet packet texture descriptors with gray-level co-occurrence matrices to generate numeric features from the images. Mala, Sadasivam, and Alagappan, 2015 specified the use of the biorthogonal wavelet and Virmani et al., 2013 used the same in addition to a variety of other wavelets.

Once features had been extracted from the images, Mala, Sadasivam, and Alagappan, 2015 and Virmani et al., 2013 narrowed down the selection of features using genetic algorithms and sequential forward searches.

To analyse the feature vectors generated Guo et al., 2009 and Mala, Sadasivam, and Alagappan, 2015 used neural networks (NNs), while Virmani et al., 2013 used support vector machines (SVMs).

2.1.3 Results

In distinguishing between HCC and cirrhotic MRI scans of rat livers, Guo et al., 2009 achieved an overall classification accuracy of 92%. However, the finer details are concerning: for the HCC images the training accuracy was 96% and testing accuracy only 83%, suggesting a high degree of overfitting. Conversely, the cirrhosis images have an accuracy in both training and testing of 100%. This suggests the classifier is too biased towards cirrhosis classification and a more effective classifier could be found if the classification boundary were shifted further towards HCC data points.

Mala, Sadasivam, and Alagappan, 2015 achieved an overall accuracy of 95% in distinguishing between fatty liver disease and cirrhosis from CT scans. The sensitivity of 96% and specificity of 94% suggest this classifier is better balanced than that of Guo et al., 2009.

Finally, Virmani et al., 2013 found an accuracy of 89% using SVMs to distinguish between healthy, cirrhotic and HCC liver ultrasound scans. They further specified a sensitivity of 90% for cirrhosis scans and 87% for HCC scans.

2.1.4 Conclusions

These examples of analysis of liver scans using machine learning show that existing techniques already give high-quality results. The three studies each achieve classification accuracies between 89-95% in distinguishing between healthy, carcinogenic, cirrhotic and fatty liver tissue images. Two use neural networks and the third uses support vector machines, showing that different methodologies can achieve similarly high results.

2.2 Hepatitis Mortality Prediction

The three studies in this category all used the same dataset from the UCI ML repository. The aim of these studies is to predict if a patient with hepatitis will die of the disease or return to health and continue living.

It is worth noting that Chen et al., 2011 incorrectly state that the purpose of the dataset is to "predict the presence or absence of hepatitis". However, this does not affect their results, since their methodologies still use the correct binary feature as the true class definitions.

2.2.1 Data

The data for these studies was taken from the UCI ML dataset named "Hepatitis Data Set"[2]. It consists of 19 predictor features (13 binary, 6 continuous) and one target feature (binary; live or die).

Faris, Aljarah, and Mirjalili, 2016 also use many other datasets, but only performance on the one UCI dataset is analysed here for comparison.

[2]http://archive.ics.uci.edu/ml/datasets/Hepatitis (page last accessed 30/10/2018)

2.2.2 Methods

Local Fisher discriminant analysis was used by Chen et al., 2011 for feature selection prior to training, while the other two papers performed no feature selection.

For classification, Chen et al., 2011 and Sartakhti, Zangooei, and Mozafari, 2012 used SVMs; the former using radial basis function (RBF) kernels and the latter using the SVM in conjunction with a simulated annealing (SA) process. Faris, Aljarah, and Mirjalili, 2016 used a multi-layer perceptron (MLP) with a single hidden layer.

2.2.3 Results

Sartakhti, Zangooei, and Mozafari, 2012 and Chen et al., 2011 both achieved near identical results, the former achieving 96% accuracy (99% sensitivity and 85% specificity) and the latter outperforming that by 1% on all three measures. Faris, Aljarah, and Mirjalili, 2016 achieved a maximum accuracy of 94% and they did not give the necessary information to calculate sensitivity or specificity metrics.

2.2.4 Conclusions

These studies show that it is possible to create classifiers using a variety of training techniques to predict mortality in hepatitis patients. Overall accuracies were between 94-97%. The SVM-based techniques here showed a slight advantage but not by a huge margin.

2.3 General liver disease detection

Of the reviewed papers there are four different classification tasks tacked by six of the studies (listed below). These all try to detect signs of one or more forms of liver disease using numeric medical data. However, the form of liver disease targeted and thus the dataset used differs between the four tasks.

The four tasks and the papers that tackle them are the detection of...

- alcoholic liver disease (erroneous, see paragraph below) Ramana, Babu, and Venkateswarlu, 2011, Olaniyi and Adnan, 2013 and Faris, Aljarah, and Mirjalili, 2016
- general liver disease (unspecified type) Sontakke, Lohokare, and Dani, 2017 and Ramana, Babu, and Venkateswarlu, 2011
- non-alcoholic fatty liver disease (NAFLD) Perveen et al., 2018
- hepatitis B virus (HBV) Mahesh, Kiruthika, and Dhilsathfathima, 2014

The first of these shall be ignored since all three papers using the dataset, UCI ML "Liver Disorders Data Set"[3], misinterpreted the data. All three sets of authors mistook a binary selector feature indicating train and test sets used by the data authors to be class labels indicating presence or absence of liver disease. The results from these analyses are therefore meaningless. Ramana, Babu, and Venkateswarlu, 2011 and Faris, Aljarah, and Mirjalili, 2016 do appear in other sections since they perform analysis on multiple datasets, whereas Olaniyi and Adnan, 2013 only used the one dataset.

2.3.1 Data

The three remaining tasks are based on three datasets, described in the following headings:

UCI Indian Liver Patient Dataset

This dataset, created by Ramana, Babu, and Venkateswarlu, 2011 and used by Sontakke, Lohokare, and Dani, 2017, is available on the UCI ML repository[4]. The dataset consists of 10 predictor features (9 continuous, 1 boolean) and a single boolean class descriptor (liver patient or non-liver patient). In the dataset version held by the UCI ML repository there are 583 records (one per patient; 416 liver patients, 167 non-liver patients). However, Sontakke, Lohokare, and Dani, 2017 use an extended version with 751 patients. This extended version also contains two further features.

[3]https://archive.ics.uci.edu/ml/datasets/liver+disorders (page last accessed 30/10/2018)

[4]https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) (page last accessed 6/11/2018)

Canadian Health System Dataset

The dataset used by Perveen et al., 2018 was gathered from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). This network contains 667,907 records recorded between 2003 and 2013. The records cover many different diseases and provide data on a range of physical attributes, vital signs, diagnoses and demographics. Perveen et al., 2018 narrow this dataset down to the 40,637 records that include all risk factors required for the study. These risk factors are based on the Adult Treatment Panel III (ATP III) guidelines and comprise gender, age, blood pressure, fasting glucose, triglycerides, high-density lipoproteins and body mass index. The authors also use other diagnoses as predictors. The aim of the research is to give a risk factor of developing non-alcoholic fatty liver disease (NAFLD). To this end, the authors define four ordinal classes and label records based on current clinical diagnostic measures.

Chronic Hepatitis B Dataset

This dataset, compiled by Mahesh, Kiruthika, and Dhilsathfathima, 2014, contains 300 records, each with 7 predictor features and 1 class label. Each predictor is a binary value denoting the presence or absence of a protein or antigen in the blood of the patient. The class label is an ordinal class value from 1 (healthy) to 4 (severe HBV).

2.3.2 Methods

For feature selection Sontakke, Lohokare, and Dani, 2017 use a ranking method before testing every subset of features ranked 1 to n where n is the number of features selected. No feature selection was performed by the other studies.

Mahesh, Kiruthika, and Dhilsathfathima, 2014 focussed on the use of generalised regression neural networks (GRNNs), Sontakke, Lohokare, and Dani, 2017 used neural networks (NNs) and support vector machines (SVMs) and Perveen et al., 2018 used decision trees. Ramana, Babu, and Venkateswarlu, 2011 tested a wide array of machine learning techniques to compare their effectiveness on medical classification tasks. They experimented with Naive Bayes (NB), decision trees (DTs), NNs, k-nearest neighbour (KNN) and sigmoid kernel SVMs (SK-SVMs).

2.3.3 Results

UCI Indian Liver Patient Dataset

The accuracy, sensitivity and specificity of the 7 different methodologies tested on the Indian Liver Patient dataset are shown in Table 2.1. The two techniques used by Sontakke, Lohokare, and Dani, 2017 show markedly lower results than the techniques used by Ramana, Babu, and Venkateswarlu, 2011. This is particularly relevant in the comparison between the neural networks used in each paper, since these used near-identical setups. One difference is that Ramana, Babu, and Venkateswarlu, 2011 used feature selection; however, this cannot account for the difference since they also quote results when using all features and still achieve accuracies above 95%. There are therefore only two documented differences in the setup that could account for the disparity: number of data samples and additional features. These are both due to Ramana, Babu, and Venkateswarlu, 2011 using the original, extended, version of the dataset which contains 168 more samples and two more features. In the feature ranking conducted on the extended dataset the third-ranked feature is "Indirect_bilirubin", which is one of the two features not present in the UCI ML repository version of the dataset used by Sontakke, Lohokare, and Dani, 2017. The authors do not state any results omitting this feature, so this seems to be the most likely reason for the difference in accuracies. The difference in numbers of samples may have some impact, though likely less than the omitted feature.

Table 2.1: Percentage accuracy, sensitivity and specificity for different algorithms applied to the UCI Indian Liver Patient dataset.

Paper	Sontakke, Lohokare, and Dani, 2017		Ramana, Babu, and Venkateswarlu, 2011				
Method	SVM	NN	NB	DT	NN	KNN	SK-SVM
Accuracy	71	73.2	95.1	96.7	97.7	97.9	96.9
Sensitivity	71.5	73.3	96.1	90	92.8	95	88.9
Specificity	88.3	87.7	94.7	98.7	99.3	98.8	99.5

Canadian Health System Dataset

The accuracy, sensitivity and specificity for each class are shown in Table 2.2. The accuracies are given as area under receiver operating characteristic curve (AUROC), since the authors did not give the required information for the accuracy measure used in the other reviewed papers. The results show a strong bias towards class 1 (healthy) making the probability of detection low for all three disease classes. In its current state it would therefore appear to be unfit for its purpose of detecting disease risk. However, the AUROC scores suggest better performance could be found (excluding class 4, whose AUROC implies near-random classification) by shifting the classification boundaries away from the healthy class, at the cost of healthy-class sensitivity.

Table 2.2: Percentage accuracy (area under receiver operating characteristic curve), sensitivity and specificity for each of the four classification categories (1=healthy to 4=very high disease risk).

Class	1	2	3	4
AUROC	74.8	63.1	73.8	50.7
Sensitivity	93.7	5	29.6	2.4
Specificity	22.3	97	94.7	99.9

Chronic Hepatitis B Dataset

The authors did not give numerical results of the performance of the classifier they designed. They gave only 5 ”sample results” without stating whether the diagnoses were correct.

2.3.4 Conclusions

While there have been a number of studies misusing a freely-available dataset, the majority of studies show promising results for detecting different forms of liver disease. Accuracies (including sensitivities and specificities) greater than 90% were achieved using a variety of machine learning techniques for the binary classification of liver disease patients from non-liver disease patients. Perveen et al., 2018 have also shown there is information available in existing electronic health records to begin the process of identifying risk of non-alcoholic fatty liver disease.

The studies that performed well enough to be clinically beneficial all depend on liver-specific markers and are therefore only viable if suspicion of liver disease is already present and specialist tests are performed. The work on Canadian health system data is a promising attempt to base classification attempts on non-specialist data but the techniques used are not yet clinically viable.

2.4 Overall Conclusions

In this chapter the use of machine learning in the diagnosis and classification of liver disease has been reviewed. The reviewed works can be grouped into three areas: Image classification / segmentation, prediction of mortality from liver disease and general liver disease diagnosis.

All three areas show good accuracies at their respective tasks when provided with specialist data relating to liver function. However, attempts to use more general health data has so far been met with limited success. This limits the scope of tests to patients already known to be at risk from liver disease, leaving the remainder of the population untouched. One study based on data made available by the Canadian health system started to look into a method of using more broadly available data but achieved low accuracies. Such an attempt may be improved if more complex machine learning tools were investigated rather than the decision trees they used.

Although mentioned in the work of Perveen et al., 2018, no attempts were made by any of the studies to track the progression of liver disease over time in patients.

Chapter 3

Prediction of the Onset of Portal Hypertension

3.1 Liver Disease in the UK

Over the past five decades liver disease has been one of the most rapidly worsening health problems in the UK. While the mortality rate of liver disease has been declining in many European countries over this time period, in the UK the mortality rate rose over 400% from 1970 to 2010, as shown in Figure 3.1. This sits in stark contrast to all other major illnesses in the UK whose mortality rates have fallen in the same given period.

One major complication in dealing with liver disease is that it exhibits no external symptoms while it progresses. Most often the first sign of the disease comes as the liver reaches cirrhosis – irreversible scarring of the organ. The most common symptoms in this case are:

- Jaundice, where the liver ceases to filter out a substance called bilirubin from the blood causing skin to turn yellow.[2]
- Abdominal swelling
- Coughing up blood, caused by portal veins in the oesophagus swelling and bursting due to increased resistance flowing through the damaged liver.

[1] http://www.euro.who.int/en/data-and-evidence/databases/european-health-for-all-database-hfa-db (last accessed 21/06/2016)

[2] http://www.nhs.uk/Conditions/Jaundice/Pages/Introduction.aspx (last accessed 02/10/2016)

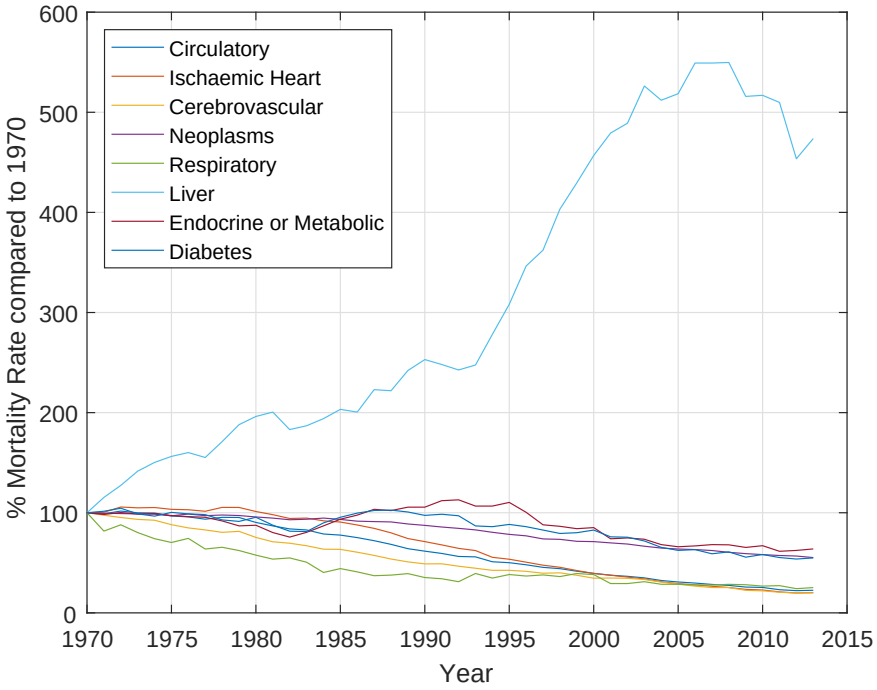


Figure 3.1: Relative changes in mortality rate for eight major diseases in the UK, normalised to 100% in 1970. Data sourced from WHO HFA database[1].

Once a liver reaches cirrhosis the only possible treatment is a liver transplant, else the disease will lead to death. However, if the disease can be picked up at an earlier stage, before external symptoms emerge, the damage can be reversed.

While there is a lot of work that can be done to tackle the major driving factor behind this trend, availability of cheap alcohol and resulting alcoholism (Anderson and Baumberg, 2006), there is also large scope for work in early detection of liver disease onset. The need for better early detection methods was outlined clearly in a meeting (and subsequent report) held in 2013 by the Foundation for Liver Research entitled "Addressing the Crisis in Liver Disease in the UK: Alcohol, Viral Hepatitis and Obesity". Out of this meeting came ten key recommendations of how to best tackle the growing problem with immediate effect. The first two of these were:

1. Strengthen detection of early liver disease and its treatment by improving the level of expertise and facilities in primary care.

2. Improve support services in the community setting for screening of high-risk patients.
(Williams et al., 2014)

Current techniques for detecting liver disease all require existing suspicion of liver damage, often due to a patient having had problems with alcoholism, obesity or hepatitis. Existing techniques also have other major drawbacks such as being invasive (liver biopsies, endoscopies) and requiring specialist, often very expensive, equipment (CT & MRI scanning, elastography) (Cadranel, Rufat, and Degos, 2000). Despite these drawbacks, the "gold standard" for liver disease diagnosis still remains as the liver biopsy, so this is the technique many studies compare results to (Bravo, Sheth, and Chopra, 2001)(Ratziu et al., 2005). Of the currently available techniques, those that best fit the requirements outlined in the 2013 meeting are the techniques based on blood tests. The first of these is the liver function test, commonly used by general practitioners in primary care with patients who are most at risk from liver disease. This test measures the levels of certain enzymes released as by-products of the fibrosis (scarring) of the liver and proteins released by the healthy function of the liver, whose levels drop in an unhealthy liver. The second blood test is the FibroTest developed by Castéra et al., 2005. However, the exact process the FibroTest uses has not been published and it is only permitted to be carried out in a few licensed laboratories making it difficult to roll out into a primary care environment.

3.2 Data

A dataset was compiled from records at Southampton General Hospital of patients who visited the hospital for a gastroendoscopy. It contains numerical results of 320,833 full blood count (FBC) blood tests taken from the 20,803 patients. Most patients will have multiple blood tests recorded in the dataset, though the number per patient varies; the median number of blood tests per patient is 8 and the maximum is 391. Each data point (FBC blood test) contains numerical readings of 13 different blood components (the 13 tests done as part of the FBC) along with the age and gender of the patient. These features and their abbreviations are listed in Table 3.1. Each patient also has an eventual diagnosis from the gastroendoscopy was (portal hypertension (PH) present / no sign of PH) and each blood test is tagged

with how many days there were between the blood test and final diagnosis. Gastroendoscopy patients were selected since PH, a consequence of around half of liver disease cases, can be accurately identified from the process. A consequence of this, however, is that although there is a control group that are PH-negative (20,089 patients vs 714 with PH) the control group cannot be assumed to be healthy. The fact that the patients have all been in need of a gastroendoscopy suggests that all patients in this dataset are in some way unwell and it is reasonable to assume that many represented illnesses will affect the patients' blood test results in some way.

Given that the majority of patients have at least two blood tests recorded, each tagged with the time prior to diagnosis, a time-series is naturally formed for each of these patients. The timeseries of each feature is shown in Figure 3.2, while a summary of the number of blood tests taken per patient can be seen in Figure 3.4. However, the gaps between blood tests are not consistent even within a single patient's results, let alone between patients. This makes direct comparisons based on time-series analysis difficult. Tests based on patient comparisons (as opposed to individual blood test comparisons) have therefore been based on data generated from the component-wise mean of each patient's blood tests. While taking the mean in this manner does remove any temporal information from the data, it allows for direct comparison between patients, even those with only one blood test result.

The distributions of each individual feature are shown in Figure 3.3. These cumulative frequency graphs, split into portal hypertension-positive and negative, give a good initial indication of which features can best split the data between the two classes. Any location on the x-axis on any of the graphs where there is a large vertical separation between the two curves demonstrates a good location for an individual linear classifier. Feature value distributions are also given in Table 3.2.

In certain tests we retained some temporal information by first segmenting the data into time cohorts based on the time prior to diagnosis of each blood test. For example, averaging over each patient's blood tests taken only in the week prior to the gastroendoscopy is likely to give more extreme values than between 1 and 2 years prior. Finally, while most tests conducted used patient comparisons as described above, in some tests we also used individual blood tests for comparison. Using individual tests has the benefit of pro-

viding much larger training and testing sets, however due to the uneven distribution of blood tests among patients any classifier trained in this way is likely to be biased towards the blood composition of those patients who have had the most blood tests.

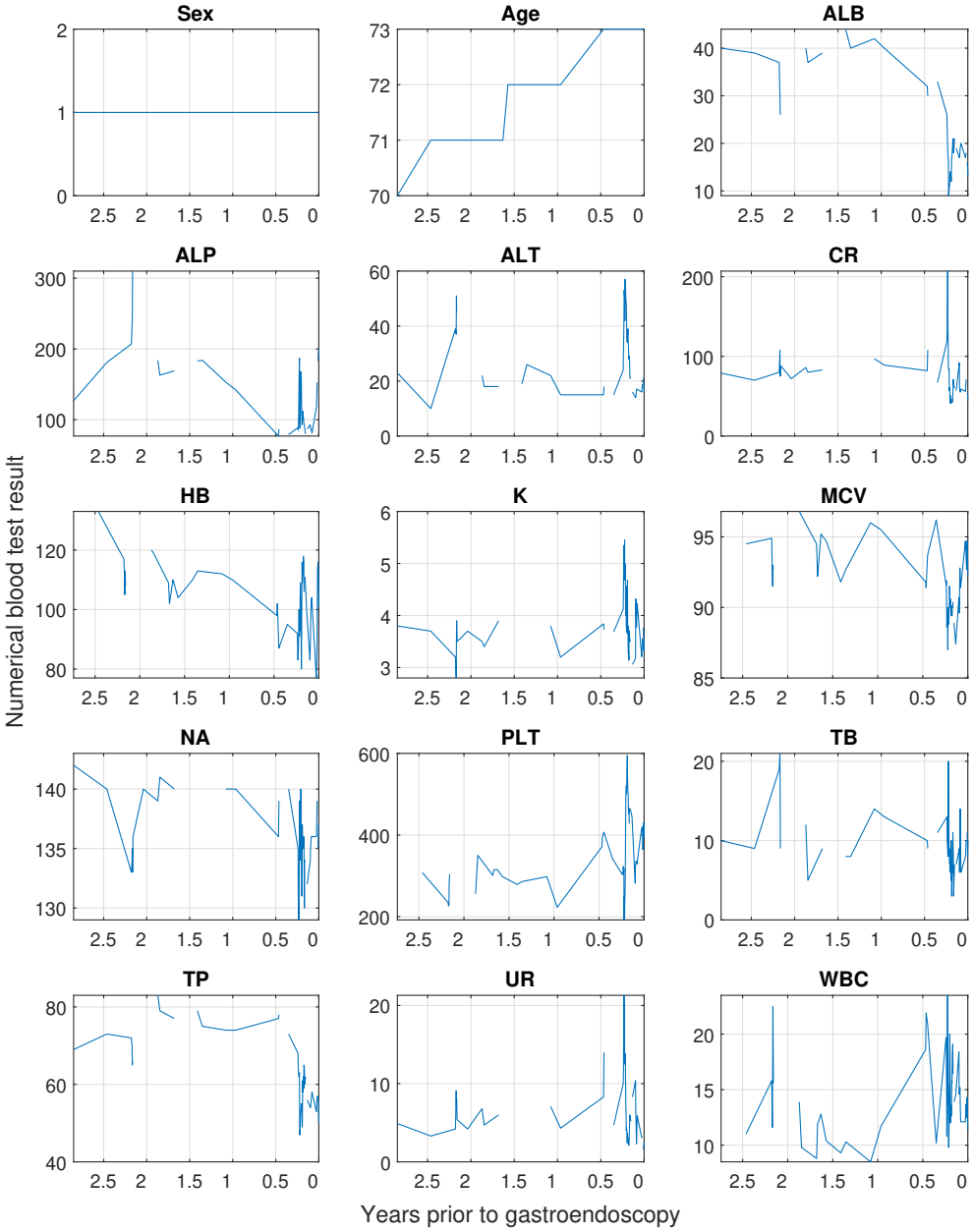


Figure 3.2: Time series of each feature taken from patient 1. The x axes show time prior to diagnosis in years. The gaps in the graphs show where that feature value is missing from a particular test.

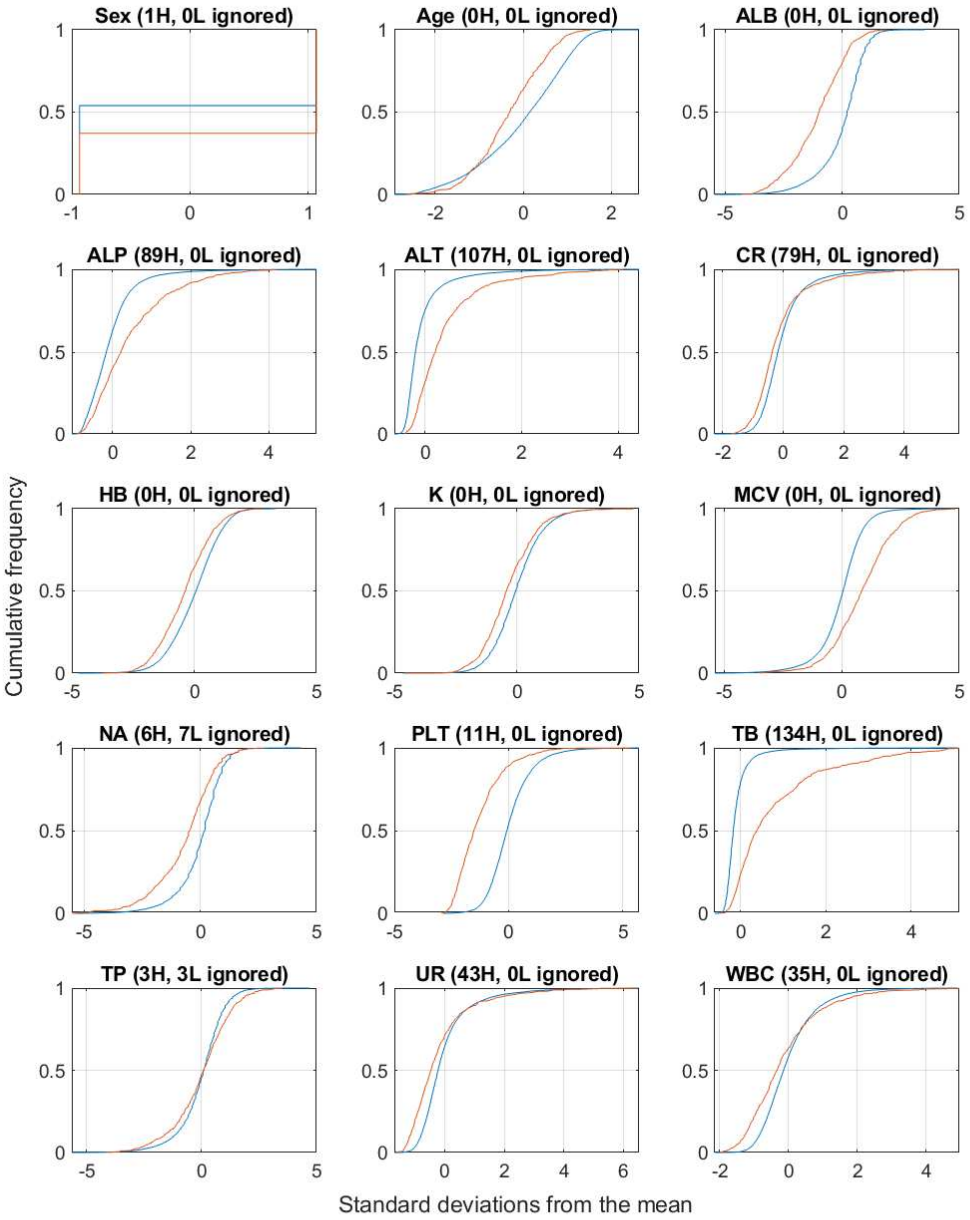


Figure 3.3: Cumulative frequency distributions of each feature, split into PH-positive (red) and PH-negative (blue). Extreme values for each feature were removed prior to splitting into PH+ and PH- and calculating CFDs so that the main bulk of the data could be visibly represented as clearly as possible. The numbers of ignored data points can be seen in the titles of each figure; #H for high values and #L for low values.

Table 3.1: Features and abbreviations

#	Abbr.	Full Name
1	Sex	Sex
2	Age	Age
3	ALB	Albumen
4	ALP	Alkaline Phosphotase
5	ALT	Alanine Aminotransferase
6	CR	Creatinine
7	HB	Haemoglobin
8	K	Potassium
9	MCV	Mean Corpuscular Volume
10	NA	Sodium
11	PLT	Platelet
12	TB	Tuberculosis
13	TP	Total Protein
14	UR	Urea
15	WBC	White Blood Cell

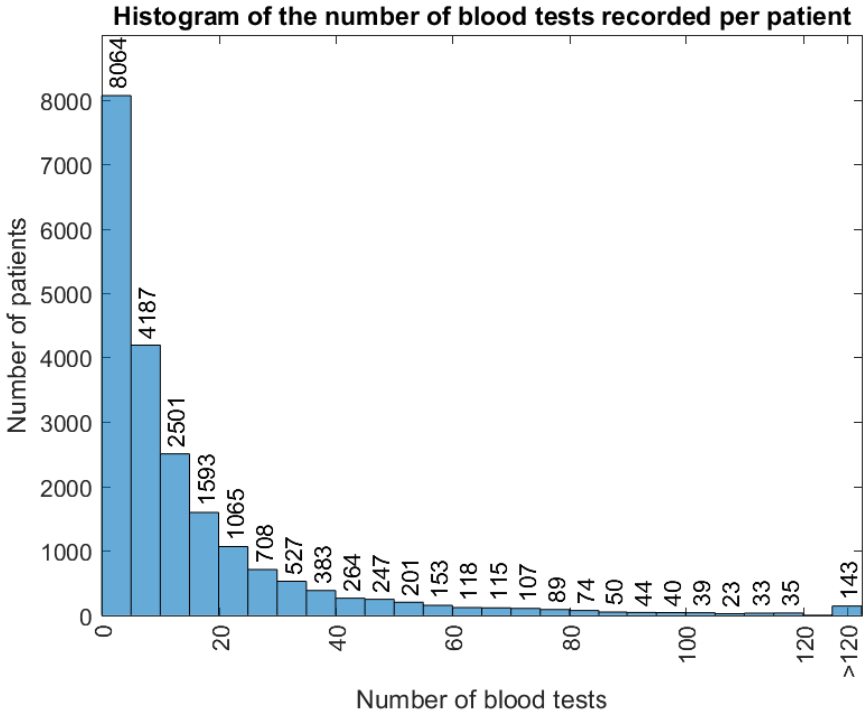


Figure 3.4: Histogram of the number of blood tests recorded per patient.

Table 3.2: Distribution of feature values across all patients

Feature	Mean	StD	Percentile						
			0	10	25	50	75	90	100
Age	65.611	16.573	6	41	56	69	78	84	109
ALB	34.446	7.1697	5	23	30	36	40	42	60
ALP	187.48	195.47	1	58	84	152	220	318	11640
ALT	36.32	94.958	1	12	16	22	34	58	6212
CR	102.41	56.435	6	63	76	90	111	146	1801
HB	120.7	22.096	10.4	91	105	122	137	149	216
K	4.0354	0.55507	1.3	3.4	3.7	4	4.34	4.7	10.2
MCV	89.056	7.283	44.7	80.6	85.2	89.2	93.1	97.2	134.6
NA	137.66	4.2384	89	132	136	138	140	142	179
PLT	273.22	125.81	1	140	198	257	328	422	1821
TB	17.046	33.814	1	6	8	11	16	24	888
TP	68.82	9.2394	16	56	64	70	75	79	141
UR	6.7357	4.9674	0.1	2.9	4	5.5	7.6	11.5	94
WBC	8.7025	5.5007	0.1	4.7	6	7.8	10.2	13.5	292.7

3.3 Methods

3.3.1 Soft Margin SVM

Soft margin support vector machines is a machine learning technique for classification of numeric data. It is described in detail in section 4.1.1

3.3.2 Logistic Regression

Logistic regression seeks to define a probability variable, $\hat{p}$ giving an estimated likelihood of a data point belonging to one of two distinct classes in order to dichotomise the dataset. First, the data, X, is normalised to have zero mean and unit standard deviation. The logistic function,

$$\hat{p} = \frac{1}{1 + e^{-X}} \tag{3.1}$$

is then used to estimate the class probability. The value of $\hat{P}$ will be between 0 and 1, where 0 is the highest likelihood of one class and 1 is the highest likelihood of the other. A threshold is defined to divide the data based on

this probability metric. This threshold is commonly set at 0.5, however it can be adjusted to identify a threshold that can better divide the data.

3.3.3 Greedy Forward Feature Selection

Greedy forward feature selection is an iterative feature selection method in which one feature is selected on each pass to be added to the feature set. The process is described in Algorithm 1.

Algorithm 1 Greedy feature selection algorithm

1: $N =$ number of features in dataset
2: Initialise $candidateFeatures = 1 : N$
3: Initialise $selectedFeatures = \{\}$
4: **while** $size(candidateFeatures) > 0 \wedge endCondition \neq true$ **do**
5: Separate dataset into train & test sets
6: Set $accuracies(f)$ to zero for all f
7: **for all** $feature$ in $candidateFeatures$ **do**
8: Train $classifier$ using $selectedFeatures$ and $feature$
9: $accuracies(feature) = test(classifier)$
10: **end for**
11: $bestAddition = \max_{feature}(accuracies(feature))$
12: Remove $bestAddition$ from $candidateFeatures$
13: Add $bestAddition$ into $selectedFeatures$
14: **end while**

3.3.4 Recursive Feature Elimination

Recursive feature elimination is another iterative method for feature selection. Unlike greedy forward feature selection which iteratively adds features, the process is started by testing a model with all possible features and eliminating one per iteration. While the process can be performed using many different models and evaluation techniques, this work focusses on the implementation based on support vector machines, as described in algorithm 2.

Algorithm 2 SVM Recursive Feature Elimination algorithm

1: N =number of features in dataset
2: Initialise $remainingFeatures = 1 : N$
3: Initialise $rejectionOrder = \{\}$
4: **while** $size(remainingFeatures) > 0 \wedge endCondition \neq true$ **do**
5: Train $classifier$ using $remainingFeatures$
6: Define $worstFeature$ as the feature with the lowest absolute weight in $classifier$
7: Remove $worstFeature$ from $remainingFeatures$
8: Add $worstFeature$ into $rejectionOrder$
9: **end while**

3.4 Results

3.4.1 Logistic Regression

A logistic regression classifier was trained on all blood tests (split into training and testing subsets). The accuracies were measured as area under receiver operating characteristic curve. The classifier achieved 92% training error and $89 \pm 1\%$ testing error (Muscat, 2015). $89 \pm 1\%$ is within the range of the FibroTest and FibroScan classifiers which achieved accuracies (when combined) of between 88-95% AUROC (Castéra et al., 2005), showing that a simple logistic regression can perform comparably with existing accepted techniques.

The logistic regression classifier was tested for comparison against an SVM classifier. Both classifiers were trained and tested on the same data segments over 10 runs of a 10-fold cross validation for a total of 100 accuracy ratings (AUROC). The results of this comparison are shown in Figure 3.5. While the logistic regression classifier is able to split data effectively, it is consistently below the performance attained by the SVM classifier.

3.4.2 Feature Selection

Feature selection is particularly important in this classification problem since the classifier needs to be as interpretable as possible. Which features (blood test components) are selected could give as much information as the classifier results and a medical professional should be able to understand why the classifier gives the result it gives.

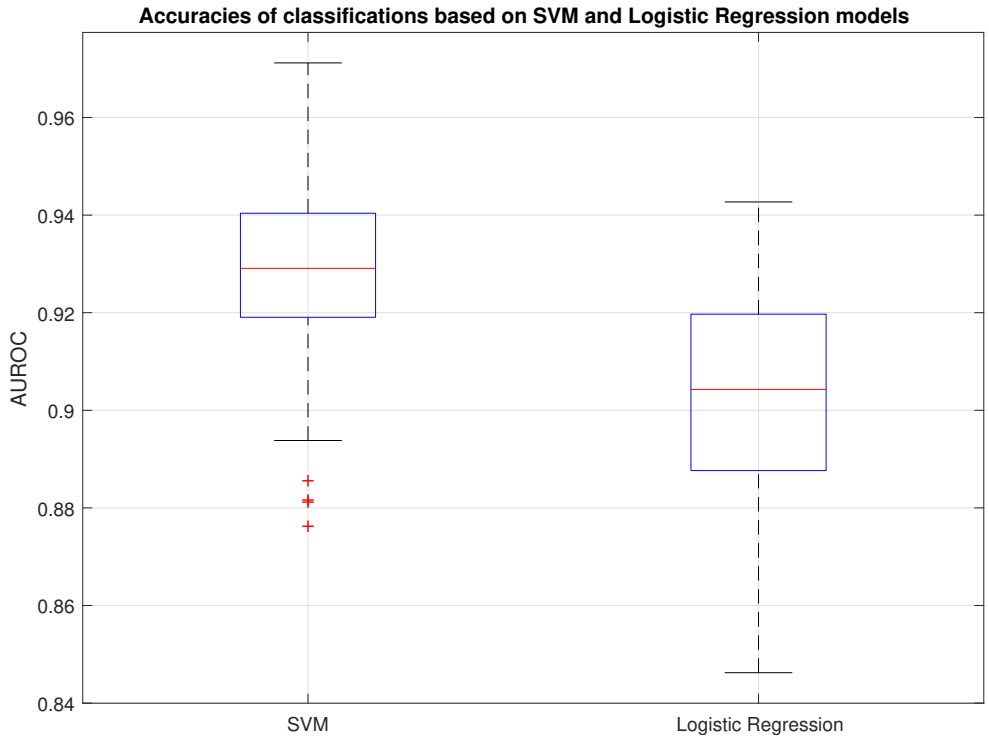


Figure 3.5: Comparison of classification accuracies between SVM and Logistic Regression classifiers.

Feature selection was performed in two groups of data; first, feature selection was performed on all data (per-patient averages) to find the best classifier overall and second, data was split into time cohorts before feature selection was performed over each time segment independently.

Selection over all data

Features TB, PLT and ALB were consistently selected as the first, second and third-most descriptive features respectively. While there was some variation in order, features TP, Age, MCV and HB were consistently selected fourth to seventh. After these seven features have been selected, the accuracy peaks at $92 \pm 0.5\%$ area under receiver operating characteristic curve (AUROC) and adding any of the remaining features causes the accuracy to dip, likely due to overfitting. These seven features are therefore taken to be the most descriptive for predicting the onset of liver disease.

The accuracy of the composite classifier as each selected feature is added

can be seen in Figure 3.6.

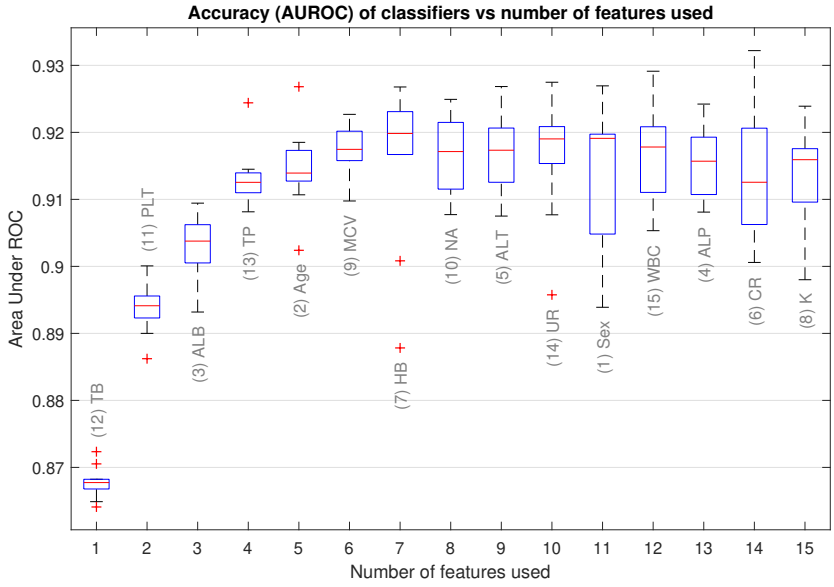


Figure 3.6: Accuracies of classifiers trained with increasing sets of selected features. Accuracies were measured using area under receiver operating characteristic curve (AUROC). Each test uses all features from all preceding tests plus the feature listed with that measurement. For example, the classifier with 4 features would use TB, PLT, ALB and TP together.

Selection over time cohorts

The reasoning behind performing feature selection on time cohorts is in seeing whether the same features selected at time of diagnosis are selected at increasing time intervals prior to diagnosis. If there are differences, this would suggest the disease affects the patient in different ways as the disease progresses, hopefully leading to a greater understanding of how the disease affects the body.

The time cohorts were defined as follows:

1. >2 years prior to diagnosis
2. 1-2 years prior to diagnosis
3. 8-365 days prior to diagnosis
4. First week prior to diagnosis

The majority of features are selected in similar positions across all cohorts, with a few exceptions. ALB is selected with high priority in cohorts 1 and 2 but not in cohorts 3 and 4, suggesting the disease progression affects that component of the blood most dramatically over a year before portal hypertension is exhibited. In contrast, TB is selected second and first in cohorts 3 and 4 respectively but given lower priority in cohorts 1 and 2, suggesting this blood component is most systematically affected close to (within a year of) diagnosis of portal hypertension.

3.4.3 Accuracy Over Time

While the ability to classify blood tests into PH/non-PH is useful, the primary aim is still to predict the onset ahead of time. To this end, a test was set up to establish how much predictive capacity there was in blood tests taken in increasing time intervals prior to diagnosis. A classifier was trained using all blood tests taken within the fortnight prior to each patient's gastroendoscopy (diagnosis date) and tested on all blood tests taken from each half-year time segment back from that date. As Figure 3.7 shows, there was already enough information in the blood tests 6 years prior to diagnosis to split the data with $75.5 \pm 1.2\%$ AUROC accuracy. This suggests that it should be possible to detect the onset of liver disease years in advance of when it would otherwise be detected in these patients.

3.4.4 Testing on Hampshire Health Record

The Hampshire Health Record is a database of over 400,000 patients' records. After the classifier was trained using the portal hypertension dataset, the weights assigned to each feature were sent to Prof. Sheron to be tested against the (previously unseen) HHR data. The classifier was run on the records of approximately 100,000 patients[3] and achieved an accuracy of 90% AUROC. This was a fraction of a percent higher than the classifier constructed by Prof. Sheron and close enough to the 92% AUROC on training data to suggest the classifier had generalised sufficiently.

[3] the remaining 300,000 were missing one or more of the features required for the classifier

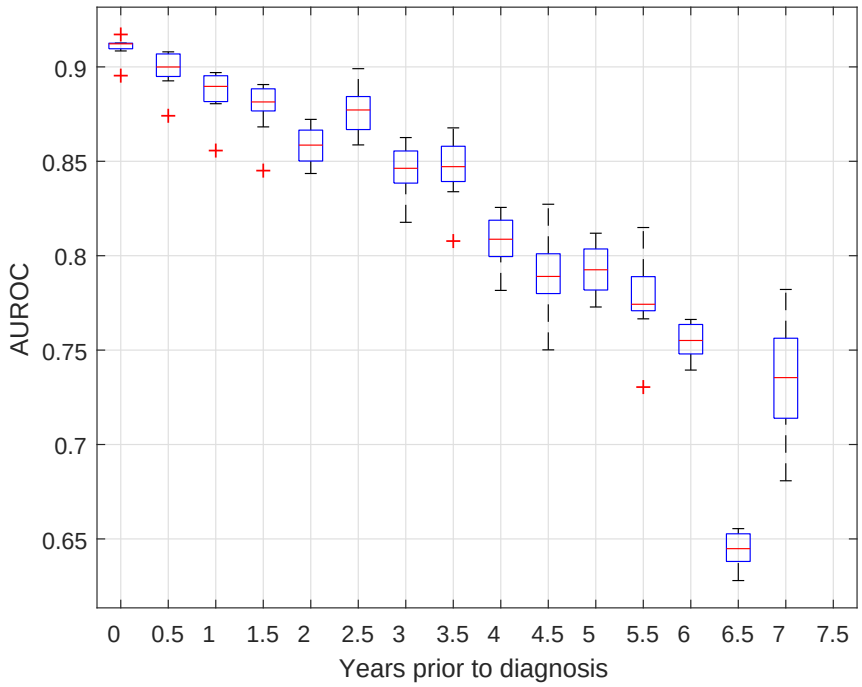


Figure 3.7: Accuracies (AUROC) of classifiers built using the 7 best selected features and tested on time segments taken from each half-year time segment back from diagnosis.

3.5 Discussion

In this chapter we have introduced a novel medical classification problem and explored solutions using a variety of machine learning techniques. Of primary importance was the ability to clarify how and why the classifier performed as it did. For this reason, feature selection was used to reduce the number of contributing factors and only linear classifiers were used to allow direct interpretation. Since medical patient data is particularly prone to systematic outliers (due to patients having heterogeneous biologies) a focus was put on identification of outliers and robust classification.

3.5.1 Feature selection

Features were selected using greedy feature selection, first from the whole dataset and then from four smaller time cohorts taken from the data. All feature selection passes found that a classifier built on the most informative

7 features achieved the peak classification accuracy and the addition of the remaining 8 features either offered no improvement or decreased accuracy, likely due to overfitting. This subset of features is particularly useful for highlighting the components of the blood test that are most affected by the onset of liver disease and thus increasing medical understanding through direct data analysis.

The feature selections on time cohorts highlighted the features whose usefulness for classification changes over time. Some features remained informative across all time cohorts, while others were better closer to diagnosis or furthest from it. This should lead to a greater understanding of how the progression of liver disease affects the blood and thus the patient.

3.5.2 Classification

Classification was performed by logistic regression and linear support vector machines. For the initial tests all blood test data was used (split into training and testing subsets) and for some later tests data was segregated based on how long prior to the patient's eventual diagnosis the blood tests were taken.

Over all data

Logistic regression achieved an accuracy of $89 \pm 1\%$ area under receiver operating characteristic curve (AUROC) and linear SVMs achieved an accuracy of $92 \pm 0.5\%$ over the same data. The baseline for these tests comes from the closest work: the FibroTest and FibroScan classifiers from Castéra et al., 2005; these baseline techniques achieved accuracies between 88% and 95% on liver disease classifications from blood tests. Our results are therefore well within the accuracy range of currently accepted techniques and crucially they use the full blood count blood test rather than specifically designed tests, such as those used by Castéra et al., 2005, giving this classifier a much wider scope for deployment in health services.

Over time cohorts

A linear SVM classifier was trained on blood tests taken from the first fortnight prior to patients' diagnoses and tested on each half-year period back from diagnosis. This showed how much information was already present in patients' blood in the years leading up to diagnosis. The results from

this were particularly promising, revealing that even 6 years prior to diagnosis the classifier trained in the fortnight of diagnosis could still separate the classes (patients with and without portal hypertension) with $75.5 \pm 1.2\%$ AUROC accuracy. This suggests that diagnosis could be made from blood tests far in advance of current methods.

Chapter 4

Outlier Rejecting Classification

4.1 Background

4.1.1 Soft Margin SVM

Definitions of terms used in this section can be found in Table 4.1.

In the standard formulation of soft margin support vector machines (SVMs) the aim in training the machine is to minimise the hinge loss function,

$$hinge(\boldsymbol{w}, \boldsymbol{x_i}, y_i) = [1 - y_i \boldsymbol{x_i}^T \boldsymbol{w}]_+ \tag{4.1}$$

The classification error for a given data pair is simply defined as 1 for incorrect class prediction and 0 for correct class prediction, as defined in (4.2).

$$err(\boldsymbol{w}, \boldsymbol{x_i}, y_i) = \begin{cases} 1 \; if \; y_i \boldsymbol{x_i}^T \boldsymbol{w} < 0 \\ 0 \; otherwise \end{cases} \tag{4.2}$$

The weights defining the SVM class boundary are found by solving:

$$\begin{aligned} &\min_{w} \frac{\beta}{2} ||\boldsymbol{w}||^2 + \sum_i [1 - y_i \boldsymbol{x_i}^T \boldsymbol{w}]_+ \\ &= \min_{w} \frac{\beta}{2} ||\boldsymbol{w}||^2 + \boldsymbol{e}^T \boldsymbol{\xi} \\ &\qquad s.t. \;\; \boldsymbol{\xi} \geqslant \boldsymbol{e} - \boldsymbol{Y} \boldsymbol{X}^T \boldsymbol{w}, \;\; \boldsymbol{\xi} \geqslant 0 \end{aligned} \tag{4.3}$$

While this formulation allows a linear SVM to find a decision boundary with overlapping classes, extreme outliers on the incorrect side of the decision boundary will still have a greater effect on the final classifier than any other

"normal" data point. In order to mitigate this effect Xu, Crammer, and Schuurmans, 2006 seek to reduce and even eliminate any influence data points perceived to be outliers have on the classifier training. They introduce a modified version of the hinge loss function,

$$\eta\text{-}hinge(\boldsymbol{w}, \boldsymbol{x_i}, y_i) = \eta_i[1 - y_i \boldsymbol{x_i}^T \boldsymbol{w}]_+ + 1 - \eta_i \tag{4.4}$$

This includes a new term, $\boldsymbol{\eta}$, which is a vector with one element for each training example. Each element $\eta_i \in [0, 1]$ defines how much effect the ith training example should have on the final classifier where $\eta_i = 1$ means the training example will have full effect as in standard hinge loss and $\eta_i = 0$ suggests the data is an outlier and causes it to have no effect on the training. This modified formulation can be solved by the following minimisation:

$$\min_{w} \min_{0 \leq \boldsymbol{\eta} \leq 1} \frac{\beta}{2} ||\boldsymbol{w}||^2 + \sum_i \eta\text{-}hinge(\boldsymbol{w}, \boldsymbol{x_i}, y_i) \tag{4.5}$$

Table 4.1: Definitions of components used in equations from Xu, Crammer, and Schuurmans, 2006

Symbol	Definition
N	Total number of data pairs
k	Total number of features
$\boldsymbol{w}$	SVM weight vector
$\boldsymbol{X}$	Input matrix consisting of N column feature vectors
$\boldsymbol{y}$	Target vector of N scalar values
$[a]_+$	$max(a, 0)$
β	Regularisation modifier; $\beta = 1$ used throughout this paper
$\boldsymbol{e}$	Vector of 1s
$\boldsymbol{x_i}$	The ith row (ith input vector) in $\boldsymbol{X}$
y_i	The ith target in $\boldsymbol{y}$
$\boldsymbol{Y}$	$diag(\boldsymbol{y})$
$err(\boldsymbol{w}, \boldsymbol{x_i}, y_i)$	Classification error using $\boldsymbol{w}$ to predict y_i from $\boldsymbol{x_i}$

4.1.2 Outlier Rejecting Regression

Robust calculation is also important for many regression problems. Most large data sets will have erroneous data (random outliers) or data points influenced by factors beyond the scope of a given experiment (systematic outliers). These two cases present subtly different challenges, though both can be tackled with similar approaches.

- *Random outliers* are typically created through random chance errors, such as mistakenly typing an incorrect value when entering records onto a computer system. This type of outlier is typically easier to detect (using nearest-neighbour clustering for example) than systematic outliers since they tend to be far more isolated in the feature space. However, for the same reason, such points can have a large impact on regressors if they are not removed prior to training - the distance of the point from the true distribution will force the regressor to shift a potentially large distance in order to accommodate it. It is therefore important, though usually not difficult, to detect and remove such outliers from the dataset prior to training.
- *Systematic outliers* occur when a bias is present in the data which is not a part of the study being conducted. For example, a doctor could frequently misdiagnose one condition as another or a study on the effects of a particular drug on patients could be skewed by multiple participants also being on a second drug. Ideally sources of systematic errors should be identified and eliminated prior to final data collection, however this is not always possible, particularly when using existing datasets for new purposes. Systematic outliers can be more difficult to identify, since they form a local distribution of their own and therefore cannot be separated by nearest-neighbour techniques alone.

Gunawardana et al., 2015 explored the use of an additional term, $\boldsymbol{\eta}$, in regression equations to mark suspected outliers and remove them from the training set. This formulation,

$$\min_{\boldsymbol{w},b} \lambda||\boldsymbol{w}||^2 + \frac{1}{(1-\mu)m}\sum_i \eta_i^k l(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b) \tag{4.6}$$

can be solved using Algorithm 3. Definitions used in (4.6) and Algorithm 3 can be found in Table 4.2. This approach will iteratively find a subset of μ

of the data points which give the lowest error, ignoring the rest of the data. This works well for eliminating random outliers but can still struggle with systematic errors, particularly if μ is set such that the number of excluded points is fewer than the number of erroneous points.

Algorithm 3 "ORR2" method from Gunawardana et al., 2015 for solving outlier rejecting regression

1: Set initial values of $\boldsymbol{w}$ and b
2: Define constants μ and λ
3: Set $k := 0$
4: $\forall_i$ calculate $l(\boldsymbol{x_i}, y_i; \boldsymbol{w}, b)$ and sort the results
5: $\forall_i$ set $\eta_i^k := \begin{cases} 0 \text{ if } l(\boldsymbol{x_i}, y_i; \boldsymbol{w}, b) \text{ was within} \\ \quad \text{the top } \mu \text{ of sorted losses} \\ 1 \text{ otherwise} \end{cases}$
6: Find $\boldsymbol{w^{k+1}}$ and b^{k+1} using (4.6)
7: Increment k by 1
8: If improvement in classification error is above a predefined threshold, return to step 4

Table 4.2: Definitions of components used in equations from Gunawardana et al., 2015

Symbol	Definition
$\boldsymbol{w}$	Weight vector
b	Bias term
λ	Regularisation parameter
μ	Fraction of the training data to be considered outliers ($0 \leq \mu \leq 1$)
m	Total number of data points in the training set
$\boldsymbol{\eta^k}$	Vector of predicted outliers on the kth iteration ($\eta_i^k \in \{0,1\};\ 0 \rightarrow outlier$)
$\boldsymbol{x_i}$	Input vector of data sample i
y_i	Target value of data sample i
$l(\boldsymbol{x_i}, y_i; \boldsymbol{w}, b)$	Generic loss function for predictor $f(\boldsymbol{x_i}) = \langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + b$ compared to true value y_i

4.2 Robust SVM Formulation

After deriving (4.5) Xu, Crammer, and Schuurmans, 2006 propose that since the equation is nonconvex a convex relaxation of the equation is required to solve it. While it is true that the equation is nonconvex, I show in this section that a simple algorithm can be derived from the equation without the need for a convex relaxation. As a nonconvex optimisation problem, this algorithm is not guaranteed to find a global optimum but its comparative simplicity to convex relaxation solutions make it an attractive approach for initial analysis of noisy data.

By substituting (4.1) into (4.4) one can observe that the *η-hinge* function can be written in terms of the original *hinge* function:

$$\eta\text{-}hinge(\boldsymbol{w}, \boldsymbol{x_i}, y_i) = \eta_i(hinge(\boldsymbol{w}, \boldsymbol{x_i}, y_i)) + 1 - \eta_i \tag{4.7}$$

This can further be substituted into (4.5) and simplified down as follows:

$$\begin{aligned} &\min_{w} \min_{0\leq\boldsymbol{\eta}\leq 1} \frac{\beta}{2}||\boldsymbol{w}||^2 + \sum_i (\eta_i[1 - y_i \boldsymbol{x_i}^T \boldsymbol{w}]_+ - \eta_i + 1) \\ &= \min_{w} \min_{0\leq\boldsymbol{\eta}\leq 1} \frac{\beta}{2}||\boldsymbol{w}||^2 + \sum_i (\eta_i([1 - y_i \boldsymbol{x_i}^T \boldsymbol{w}]_+ - 1) + 1) \\ &= \min_{w} \min_{0\leq\boldsymbol{\eta}\leq 1} \frac{\beta}{2}||\boldsymbol{w}||^2 + \sum_i \eta_i([1 - y_i \boldsymbol{x_i}^T \boldsymbol{w}]_+ - 1) + N \end{aligned} \tag{4.8}$$

The sum term can then be transformed into a vector equation by defining $\boldsymbol{\xi}$ as used in (4.3)

$$\begin{aligned} &\min_{w} \min_{0\leq\boldsymbol{\eta}\leq 1} \frac{\beta}{2}||\boldsymbol{w}||^2 + \sum_i (\eta_i(\xi_i - 1)) + N \\ &\qquad\qquad s.t.\ \forall_i(\xi_i \geq 1 - y_i \boldsymbol{x_i}^T \boldsymbol{w},\ \xi_i \geq 0) \\ &= \min_{w} \min_{0\leq\boldsymbol{\eta}\leq 1} \frac{\beta}{2}||\boldsymbol{w}||^2 + \boldsymbol{\eta}^T(\boldsymbol{\xi} - \boldsymbol{e}) + N \\ &\qquad\qquad s.t.\ \boldsymbol{\xi} \geq \boldsymbol{e} - \boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w},\ \boldsymbol{\xi} \geq 0 \end{aligned} \tag{4.9}$$

By defining $\boldsymbol{\alpha} = \boldsymbol{\xi} - \boldsymbol{e}$ the following alternating minimisation problem can be found:

$$\begin{aligned} &define:\ \boldsymbol{\alpha} = \boldsymbol{\xi} - \boldsymbol{e} \\ &\min_{w} \frac{\beta}{2}||\boldsymbol{w}||^2 + \boldsymbol{\eta}^T\boldsymbol{\xi} \qquad s.t.\ \boldsymbol{\xi} \geq \boldsymbol{e} - \boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w},\ \boldsymbol{\xi} \geq 0 \\ &\min_{0\leq\boldsymbol{\eta}\leq 1} \boldsymbol{\eta}^T\boldsymbol{\alpha} \qquad\qquad s.t.\ \boldsymbol{\alpha} \geq -\boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w},\ \boldsymbol{\alpha} \geq -1 \end{aligned} \tag{4.10}$$

The first half of this minimisation can be seen as solving the hinge loss soft margin SVM for all data pairs whose corresponding value of η is 1 (not an outlier). The second half of the minimisation, finding $\boldsymbol{\eta}$, can be achieved through calculating the error, (4.2), which adapts the alternating minimisation to

$$\begin{gathered}\min_{w} \frac{\beta}{2}||\boldsymbol{w}||^2 + \boldsymbol{\eta}^T\boldsymbol{\xi} \;\; s.t. \; \boldsymbol{\xi} \geq \boldsymbol{e} - \boldsymbol{Y}\boldsymbol{X}^T\boldsymbol{w}, \; \boldsymbol{\xi} \geq 0 \\ \forall_i(\eta_i := 1 - err(\boldsymbol{w}, \boldsymbol{x_i}, y_i))\end{gathered} \tag{4.11}$$

This final formulation is easily computable using Algorithm 4.

Algorithm 4 Iterative method for computing the *η-hinge* SVM, (4.11)

1: Initialise $\boldsymbol{\eta} := \boldsymbol{e}$
2: **repeat**
3: Solve the hinge loss soft margin SVM for all data pairs $(\boldsymbol{x_i}, y_i)$ where $\eta_i = 1$
4: **for all** i **do**
5: Predict $\hat{y}_i$ from $\boldsymbol{x_i}$ using the computed SVM
6: Set $prev\text{-}\eta_i := \eta_i$
7: Set $\eta_i := \begin{cases} 1 \text{ if } \hat{y}_i = y_i \\ 0 \text{ otherwise} \end{cases}$
8: **end for**
9: **until** $\forall_i(prev\text{-}\eta_i = \eta_i)$

4.2.1 Class Balancing

The primary issue with this formulation is revealed when there is a large difference between the sizes of the two classes. The algorithm will often end up marking all of the minority class as outliers and placing the decision boundary arbitrarily far from the remaining class such that it achieves 100% accuracy on the majority class. To avoid this behaviour an additional weight, φ_i was added to the equation and set such that the sum of weights of each class was equal. This ensured classification boundaries were drawn more centrally between the classes rather than drifting towards (and beyond) the minority class over successive iterations.

$$
\begin{aligned}
define: N_A &= \text{number of data pairs belonging to class A} \\
N_B &= \text{number of data pairs belonging to class B} \\
class(i) &= \text{the class that data pair i belong to} \\
\varphi_i &= \frac{max(N_A, N_B)}{class(i)} \\
\min_{w} \min_{0 \leq \boldsymbol{\eta} \leq 1} & \frac{\beta}{2}||\boldsymbol{w}||^2 + \sum_i \varphi_i(\eta_i[1 - y_i \boldsymbol{x_i}^T \boldsymbol{w}]_+ - \eta_i + 1)
\end{aligned}
\tag{4.12}
$$

By following through the process taken from (4.8) to (4.11) but starting from (4.12) rather than (4.8), the following alternating minimisation problem is found:

$$
\begin{gathered}
\min_{w} \frac{\beta}{2}||\boldsymbol{w}||^2 + \boldsymbol{\eta}^T \boldsymbol{\Phi} \boldsymbol{\xi} \;\; s.t. \; \boldsymbol{\xi} \geq \boldsymbol{e} - \boldsymbol{Y} \boldsymbol{X}^T \boldsymbol{w}, \; \boldsymbol{\xi} \geq 0 \\
\forall_i (\eta_i := 1 - err(\boldsymbol{w}, \boldsymbol{x_i}, y_i))
\end{gathered}
\tag{4.13}
$$

4.3 Robust SVM Performance

The robust SVM was first tested on a small artificial dataset from Xu, Crammer, and Schuurmans, 2006 to verify it was behaving as expected. This dataset is made up of 20 samples drawn from two distributions with a final 10 data points sampled from a ring as outliers for a total of 50 two-dimensional data points. The two distributions have means of $\mu_1 = (3, -3)$ and $\mu_2 = (-3, 3)$ and both share the same covariance matrix $\Sigma = \begin{pmatrix} 20 & 16 \\ 16 & 20 \end{pmatrix}$. The ring the outliers are drawn from is centred around the origin with inner radius of 15 and outer radius of 16. The 10 data points drawn from this ring are randomly assigned to either of the classes with equal probability.

This artificial dataset was used to train first a standard soft margin SVM and then a robust (outlier rejecting) SVM. The results of these two training methods are shown in Fig. 4.1. Despite the classification boundary being very similar in both cases, the ORSVM is visibly more confident in the boundary it finds after rejecting the highlighted outliers. This should lead to a more accurate representation of class boundary since it is unbiased from extreme erroneous data.

While the ORSVM algorithm can be seen to work on small datasets such as that described above, it has so far shown no significant improvement to

classification accuracies in large datasets. It is currently unknown if this is more affected by the total number of data points or total number of features used. Clearly, it is also unlikely to give any benefit for datasets with few outliers.

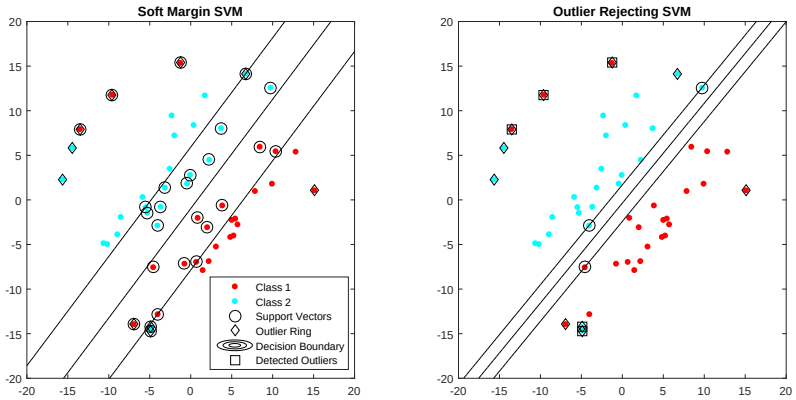


Figure 4.1: Comparison of class and support vector boundaries found by a Soft Margin SVM and Outlier Rejecting SVM on identical 2-dimensional artificial datasets with outliers.

4.4 Discussion

A robust SVM was implemented based on the work of Xu, Crammer, and Schuurmans, 2006 that identifies outliers while training and returns an SVM trained on the remainder of the data along with the list of predicted outliers. An extra term, φ_i, was added to the robust SVM equation to balance the outlier rejection in cases where the positive and negative classes are not the same size.

The final formulation worked well on small-scale, artificial datasets. However, the predicted increase in accuracy from this method did not scale up to larger datasets like the portal hypertension dataset used in Chapter 3. Combining the methodology with a cluster-based outlier identification technique could lead to improvements or insights as to why the algorithm in its current form does not scale up well.

Despite not providing an increase in accuracy, the list of outliers identified by multiple passes of the algorithm should provide some insights into larger

datasets (such as the portal hypertension dataset), though this is a property of outlier identification in general and not exclusive to this methodology.

Chapter 5

Conclusions & Future Work

5.1 Conclusions

5.1.1 Liver Disease Classification

Prior Work

Chapters 1 and 2 investigated the work that has been done to date in applying machine learning to medicine in general and liver disease, respectively. While there were some examples found of the use of general electronic medical records (EMRs) in medical diagnoses, almost all work in liver disease diagnosis used information specific to liver function. As discussed in chapter 3 this presents the issue of narrowing down the scope of the tests to be used only in cases where there is existing reason to suspect liver disease. A system was required which can identify liver disease in patients who would otherwise go untested.

Analysis of Portal Hypertension Data

To build a system which can detect early signs of liver disease with non-liver specific information a dataset collected from inpatients at Southampton General Hospital was analysed. The data consisted of full blood count test results (a very common type of test frequently conducted and recorded in general practices and hospitals), along with basic information like patient sex and age. The dataset also contained diagnosis information for portal hypertension, a reliable symptom of liver disease. The names of the blood components measured were masked (replaced with labels A-O) in order to

avoid any bias towards previous research in the creation of the classifier. The key of letter label to feature name was retained by the liver specialist who gathered the data so the eventual classifier could be verified and learned from.

Feature selection was performed on the data to identify the best correlated components of the full blood count data with the presence of portal hypertension. Of the 15 features, 3 were found to carry the majority of the information ($> 90\%$ classification accuracy) and 7 features were found to perform best ($92 \pm 0.5\%$ classification accuracy). Adding further features decreased classification accuracy suggesting they gave no relevant independent information and instead lead to overfitting. The features identified were confirmed by the liver specialist to relevant in our current understanding of liver disease progression.

Classification methods were explored for identifying markers in the full blood count data that could predict the onset of portal hypertension. Linear support vector machines were the most thoroughly tested classification tool, since they can create a robust classification boundary while remaining fully interpretable. The final SVM classifier achieved a performance of $92 \pm 0.5\%$ area under receiver operating characteristic curve (AUROC), a minor improvement on the logistic regression classifier designed by the liver specialist. The classifier was also tested on a previously-unseen dataset of around 100,000 patients in Hampshire, UK. On this data an accuracy of 90% was achieved, again marginally outperforming the classifier designed by the liver specialist.

While the majority of the research was conducted on patient data averaged over time (to create one data point per patient) the final investigations looked into the temporal information content of the portal hypertension dataset. First, the dataset was split into four time cohorts based on how long prior to diagnosis the blood test was conducted (thresholds at 1 week, 1 year and 2 years). Feature selection was performed on data from these cohorts to see if the most informative features changed over time. Features 3 and 12 varied most over time, with 12 being most important for very early detection and 3 most important nearer diagnosis. The dataset was further divided into half-year cohorts and tested with a classifier trained on the fortnight immediately prior to diagnosis. This revealed a downward trend

the further back from diagnosis date the data went, as would be expected. However, most importantly, this demonstrated that a classification accuracy of over 75% AUROC could be achieved six years prior to the date patients were eventually diagnosed. This shows that there is a significant amount of information available early enough in the disease progression to halt the damage and prevent liver cirrhosis.

5.1.2 Outlier Rejecting Classification

In chapter 4 techniques for automatically identifying and eliminating outliers were explored. A support vector machine-based technique was developed drawing from the work introduced by Xu, Crammer, and Schuurmans, 2006 and Gunawardana et al., 2015.

The outlier-rejecting support vector machine (ORSVM) was tested on a small artificial dataset with two bivariate Gaussian classes and a small number of erroneous outliers. On this dataset the power of such a system was demonstrated as the classifier removed all outliers prior to final training, resulting in a much tighter (more certain) classification margin.

The technique was then tested on larger scale data including the portal hypertension dataset from chapter 3. On this data an issue with identifying outliers in datasets with unbalanced class sizes was found. If one class was sufficiently small, the algorithm could assume all of that class was erroneous and ignore it. This brought accuracy to 100% but made any classifications meaningless. To mitigate this a further term was added into the formulated equation to balance the proportion of outliers identified from each class. This stopped the classifier eliminating the minority class. However, it still showed little (if any) improvement over standard soft-margin SVMs when applied to large-scale datasets.

In its current form the outlier-rejecting support vector machine performs well wherever there is significant separation between accurate and erroneous data but in situations where erroneous data is close to the true distribution of a class ORSVM shows no advantage over existing techniques like the soft-margin SVM.

5.2 Future Work

5.2.1 Outlier Rejecting Classification

Despite the lacking benefit demonstrated by the current ORSVM formulation, the need for an accurate method of outlier rejection remains. One of the situations ORSVM does not tackle is outlier clusters that fall the 'correct' side of the classification boundary. A potential solution to this may be found in drawing on data clustering techniques and attempting to integrate them into classification methods. These may include SVMs, however other methods should be investigated too.

5.2.2 Portal Hypertension Dataset

RETAIN

Section 1.2 introduced the issue of accuracy vs interpretability. All of the methods used thenceforth have been based on naturally interpretable methods adjusted to be as accurate as possible. An alternative solution would be to take a naturally highly accurate method and adjust it to be interpretable. This is the approach taken by Choi et al., 2016. They introduce a recurrent neural network model (REverse Time AttentIoN model - RETAIN) along with methods to analyse the model after training to understand how the model functions. This was used for the prediction of clinical diagnoses from health record data, a near identical application to that discussed in Chapter 3. Application and development of the RETAIN model to liver disease prediction has two primary appeals:

- Recurrent neural networks (such as RETAIN) have far more scope for learning complex patterns, in comparison to the linear models researched in this thesis

- A lot of information is lost when health record values are averaged over time. RETAIN was designed with time as one of the most important pieces of information, so using this form of analysis should tap into a very useful extra piece of data.

New SGH dataset

Since the completion of this research a new dataset has become available to Prof. Sheron compiled from in-patient data at Southampton General Hospital. The new dataset has multiple benefits over the portal hypertension dataset used in this research:

- It is not restricted to gastroendoscopy patients, so it should better represent the general population.
- It is an order of magnitude larger, both in number of patients and number of records, so there is greater scope for looking at finer details without overfitting.
- It contains additional fields that the PH dataset lacked, such as diagnosis codes (what was wrong with the patient when they were admitted). These codes would be particularly helpful in identifying and eliminating systematic outliers caused by other common conditions and, as a result, could assist in training classifiers to automatically detect such groups of cases.

5.2.3 Natural Language Processing on Liver Biopsy Transcripts

The current process (outlined below) for conducting and analysing liver biopsies in Southampton General Hospital involves a significant administrative overhead. Since the biopsy transcripts generated in stage 2 are in a common format, a suggested task from Prof. Sheron is to automate stage 3 (generating the fibrosis score from the text).

1. Doctor speaks notes into a dictaphone as they perform the biopsy.
2. Recording from dictaphone is sent to an external company to be transcribed and returned.
3. Transcript is read by the doctor or a colleague/student and graded from 0-4 (fibrosis score).
4. Fibrosis score is used by the doctor to make a diagnosis.

Data

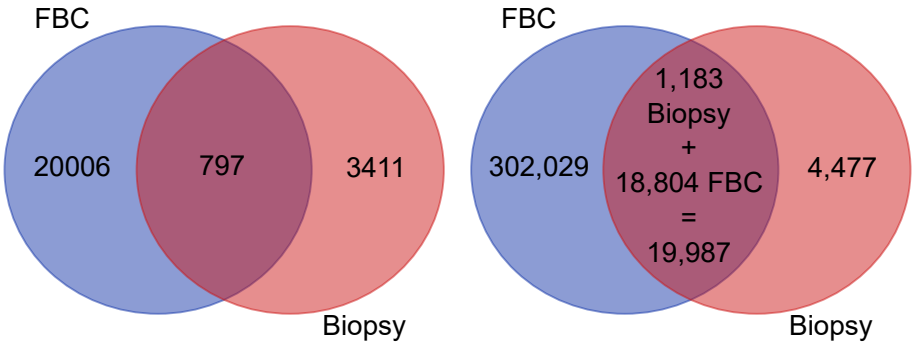
A dataset of information taken from 5,659 liver biopsies was provided by Prof. Sheron. There are 41 fields in the dataset in total, of which 7 are of

particular interest. These 7 are described in Table 5.1.

Table 5.1: Descriptions of important fields in the liver biopsy dataset

Field	Data Type	Data Description
Bencode	Numeric	Patient code transformed for anonymity. This code will be consistent for any given patient over the multiple datasets sourced by Prof. Sheron allowing for tests to be performed involving multiple datasets
Cirrhosis	Boolean	Marker indicating whether or not the analysed liver was cirrhotic
Diagnosis	Text	
Datebiopsy	Date	Date the liver biopsy was taken
Fibrosisscore	Integer (0-4)	A measure of how badly scarred the liver was, from 0 indicating a healthy liver, through to a 4 indicating a cirrhotic liver
Reportsummary	Text	Transcript of notes taken during the liver biopsy
SNOMEDcodes	Text	A list of medical codes fitting the condition of the liver

The liver biopsy dataset overlaps with the portal hypertension dataset (Section 3.2) on 797 patients. These patients account for 1,183 of the 5,659 biopsy records and 18,804 of the 320,833 blood tests (see Figure 5.1).



(a) Patients (b) Records

Figure 5.1: Venn diagrams showing how many patients are shared between the portal hypertension dataset and the biopsy dataset (a) and how many records belong to these shared patients (b)

Bibliography

Anderson, Peter and Ben Baumberg (2006). *Alcohol in Europe*. London: Institute of Alcohol Studies.

Bravo, Arturo A, Sunil G Sheth, and Sanjiv Chopra (2001). "Liver biopsy". In: *New England Journal of Medicine* 344.7, pp. 495–500.

Cadranel, Jean-François, Pierre Rufat, and Françoise Degos (2000). "Practices of liver biopsy in France: results of a prospective nationwide survey". In: *Hepatology* 32.3, pp. 477–481.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad (2015). "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission". In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 1721–1730.

Castéra, Laurent, Julien Vergniol, Juliette Foucher, Brigitte Le Bail, Elise Chanteloup, Maud Haaser, Monique Darriet, Patrice Couzigou, and Victor de Lédinghen (2005). "Prospective comparison of transient elastography, Fibrotest, APRI, and liver biopsy for the assessment of fibrosis in chronic hepatitis C". In: *Gastroenterology* 128.2, pp. 343–350.

Chen, Hui-Ling, Da-You Liu, Bo Yang, Jie Liu, and Gang Wang (2011). "A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis". In: *Expert Systems with Applications* 38.9, pp. 11796–11803.

Choi, Edward, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart (2016). "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism". In: *Advances in Neural Information Processing Systems*, pp. 3504–3512.

Faris, Hossam, Ibrahim Aljarah, and Seyedali Mirjalili (2016). "Training feed-forward neural networks using multi-verse optimizer for binary classification problems". In: *Applied Intelligence* 45.2, pp. 322–332.

Gunawardana, Yawwani, Shuhei Fujiwara, Akiko Takeda, Jeongmin Woo, Christopher Woelk, and Mahesan Niranjan (2015). "Outlier detection at the transcriptome-proteome interface". In: *Bioinformatics*, btv182.

Guo, Dongmei, Tianshuang Qiu, Jie Bian, Wei Kang, and Li Zhang (2009). "A computer-aided diagnostic system to discriminate SPIO-enhanced magnetic resonance hepatocellular carcinoma by a neural network classifier". In: *Computerized Medical Imaging and Graphics* 33.8, pp. 588–592.

Havaei, Mohammad, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle (2017). "Brain tumor segmentation with deep neural networks". In: *Medical image analysis* 35, pp. 18–31.

Leung, Michael KK, Andrew Delong, Babak Alipanahi, and Brendan J Frey (2016). "Machine learning in genomic medicine: a review of computational problems and data sets". In: *Proceedings of the IEEE* 104.1, pp. 176–197.

Mahesh, C, K Kiruthika, and M Dhilsathfathima (2014). "Diagnosing hepatitis B using artificial neural network based expert system". In: *International Conference on Information Communication and Embedded Systems (ICICES2014)*. IEEE, pp. 1–7.

Mahmud, Nadim, Jonah Cohen, Kleovoulos Tsourides, and Tyler M Berzin (2015). "Computer vision and augmented reality in gastrointestinal endoscopy". In: *Gastroenterology report* 3.3, pp. 179–184.

Mala, K, V Sadasivam, and S Alagappan (2015). "Neural network based texture analysis of CT images for fatty and cirrhosis liver classification". In: *Applied Soft Computing* 32, pp. 80–86.

Muscat, Anne-Marie (2015). "Data Driven Modelling for Predicting the Onset of Liver Disease". MA thesis. The University of Southampton.

Olaniyi, Ebenezer Obaloluwa and Khashman Adnan (2013). "Liver Disease Diagnosis Based on Neural Networks". In: *Advances in Computational Intelligence*.

Onu, Charles C, Innocent Udeogu, Eyenimi Ndiomu, Urbain Kengni, Doina Precup, Guilherme M Sant'anna, Edward Alikor, and Peace Opara (2017). "Ubenwa: Cry-based Diagnosis of Birth Asphyxia". In: *arXiv preprint arXiv:1711.06405*.

Patel, Ankeeta R and Maulin M Joshi (2013). "Heart diseases diagnosis using neural network". In: *Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on*. IEEE, pp. 1–5.

Perveen, Sajida, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi (2018). "A Systematic Machine Learning Based Approach for the Diagnosis of Non-Alcoholic Fatty Liver Disease Risk and Progression". In: *Scientific reports* 8.1, p. 2112.

Pons, Ewoud, Loes MM Braun, MG Myriam Hunink, and Jan A Kors (2016). "Natural language processing in radiology: a systematic review". In: *Radiology* 279.2, pp. 329–343.

Ramana, Bendi Venkata, M Surendra Prasad Babu, and NB Venkateswarlu (2011). "A critical study of selected classification algorithms for liver disease diagnosis". In: *International Journal of Database Management Systems* 3.2, pp. 101–114.

Ratziu, Vlad, Frédéric Charlotte, Agnès Heurtier, Sophie Gombert, Philippe Giral, Eric Bruckert, André Grimaldi, Frédérique Capron, Thierry Poynard, LIDO Study Group, et al. (2005). "Sampling variability of liver biopsy in nonalcoholic fatty liver disease". In: *Gastroenterology* 128.7, pp. 1898–1906.

Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga (2007). "A review of feature selection techniques in bioinformatics". In: *bioinformatics* 23.19, pp. 2507–2517.

Sartakhti, Javad Salimi, Mohammad Hossein Zangooei, and Kourosh Mozafari (2012). "Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA)". In: *Computer methods and programs in biomedicine* 108.2, pp. 570–579.

Shameer, Khader, Kipp W Johnson, Alexandre Yahi, Riccardo Miotto, LI Li, Doran Ricks, Jebakumar Jebakaran, PATRICIA KOVATCH, Partho P Sengupta, SENGUPTA GELIJNS, et al. (2017). "Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai Heart Failure Cohort". In: *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*. World Scientific, pp. 276–287.

Sontakke, Sumedh, Jay Lohokare, and Reshul Dani (2017). "Diagnosis of liver diseases using machine learning". In: *Emerging Trends & Innovation in ICT (ICEI), 2017 International Conference on*. IEEE, pp. 129–133.

Virmani, Jitendra, Vinod Kumar, Naveen Kalra, and Niranjan Khandelwal (2013). "SVM-based characterization of liver ultrasound images using wavelet packet texture descriptors". In: *Journal of digital imaging* 26.3, pp. 530–543.

Williams, Roger, Richard Aspinall, Mark Bellis, Ginette Camps-Walsh, Matthew Cramp, Anil Dhawan, James Ferguson, Dan Forton, Graham Foster, Ian Gilmore, et al. (2014). "Addressing liver disease in the UK: a blueprint for attaining excellence in health care and reducing premature mortality from lifestyle issues of excess consumption of alcohol, obesity, and viral hepatitis". In: *The Lancet* 384.9958, pp. 1953–1997.

Xu, Linli, Koby Crammer, and Dale Schuurmans (2006). "Robust support vector machine training via convex outlier ablation". In: *AAAI*. Vol. 6, pp. 536–542.