The University of Sheffield

Department of Automatic Control
and Systems Engineering

Faculty of Engineering

Automatic
Control &
Systems
Engineering.

Nonlinear Parametric and Neural Network Modelling for Medical Image Classification

by

Carlos Beltran Perez

A thesis submitted in partial fulfilment
of requirements for the degree of

Doctor of Philosophy

Sheffield, England, United Kingdom November 2018

To my loving parents Manuel Luis and Luz.

Acknowledgements

Thanks to my mother Lucha for her infinite patience and love, to my girlfriend Sole for her unconditional love and support and to my brother Luis, my sister Claudia and my brother in law Carlos for being by my side no matter the distance.

Thanks to Dr Hua-Liang Wei, for whom I was proudly supervised and whom I consider the best support and guidance I could have in my Doctoral studies.

Thanks to my friends Roberto Ayala, Carlos Luna, Adrian Rubio, Marco Galindo, Jozra Garrido, Myriam Gomez, Sergio Rodriguez, Ricardo Viteri, Paulina Gonzalez, Mauro Cruz, Jose Avalos, Karla Valdez, Emilio Pliego, Karla Tun, Hector Balboa, Duman Furrer, Manuel Valera, Olivia Espinosa and Mike Hoylland for bringing fun and support to my life.

Thanks to my colleagues in ACSE Rajintha Gunawardena, Mohammed Hazim, Antonio Penuelas, Alex Vidal, Malu Davila and Rafa Colas for your support and friendship.

Thanks to my PhD Examiners for the time invested in reading and evaluating my dissertation.

Thanks to the University of Sheffield and the Department of Automatic Control and System Engineering for giving me the chance to study at a world-class university. Special thanks to Matthew and Renata.

Thanks to the entire ACSE community for being part of my personal and professional growth process.

Thanks to CONACyT, for the financial support and for allowing me to study abroad in an excellent British institution.

Abstract

System identification and artificial neural networks (ANN) are families of algorithms used in systems engineering and machine learning respectively that use structure detection and learning strategies to build models of complex systems by taking advantage of input-output type data. These models play an essential role in science and engineering because they fill the gap in those cases where we know the input-output behaviour of a system, but there is not a mathematical model to understand and predict its changes in future or even prevent threats. In this context, the nonlinear approximation of systems is nowadays very popular since it better describes complex instances. On the other hand, digital image processing is an area of systems engineering that is expanding the analysis dimension level in a variety of real-life problems while it is becoming more attractive and affordable over time. Medicine has made the most of it by supporting important human decision-making processes through computer-aided diagnosis (CAD) systems.

This thesis presents three different frameworks for breast cancer detection, with approaches ranging from nonlinear system identification, nonlinear system identification coupled with simple neural networks, to multilayer neural networks. In particular, the nonlinear system identification approaches termed the Nonlinear AutoRegressive with eXogenous inputs (NARX) model and the MultiScales Radial Basis Function (MSRBF) neural networks appear for the first time in image

processing. Along with the above contributions takes place the presentation of the Multilayer-Fuzzy Extreme Learning Machine (ML-FELM) neural network for faster training and more accurate image classification.

A central research aim is to take advantage of nonlinear system identification and multilayer neural networks to enhance the feature extraction process, while the classification in CAD systems is bolstered. In the case of multilayer neural networks, the extraction is carried throughout stacked autoencoders, a bottleneck network architecture that promotes a data transformation between layers. In the case of nonlinear system identification, the goal is to add flexible models capable of capturing distinctive features from digital images that might be shortly recognised by simpler approaches. The purpose of detecting nonlinearities in digital images is complementary to that of linear models since the goal is to extract features in greater depth, in which both linear and nonlinear elements can be captured. This aim is relevant because, accordingly to previous work cited in the first chapter, not all spatial relationships existing in digital images can be explained appropriately with linear dependencies.

Experimental results show that the methodologies based on system identification produced reliable images models with customised mathematical structure. The models came to include nonlinearities in different proportions, depending upon the case under examination. The information about nonlinearity and model structure was used as part of the whole image model. It was found that, in some instances, the models from different clinical classes in the breast cancer detection problem presented a particular structure. For example, NARX models of the malignant class showed higher non-linearity percentage and depended more on exogenous inputs compared to other classes.

Regarding classification performance, comparisons of the three new CAD systems with existing methods had variable results. As for the NARX model, its performance was superior in three cases but was overcame in two. However, the comparison must be taken with caution since different databases were used. The MSRBF model was better in 5 out of 6 cases and had superior specificity in all instances, overcoming in 3.5% the closest model in this line. The ML-FELM model was the best in 6 out of 6 cases, although it was defeated in accuracy by 0.6% in one case and specificity in 0.22% in another one.

Contents

Acknowledgements **iii**

Nomenclature **xxi**

1 Introduction **1**

1.1 Background 1

1.2 Motivation 3

1.3 Aims and objectives 4

1.3.1 General objective 4

1.3.2 Specific aims 5

1.4 Thesis overview 5

1.5 Main contributions 7

1.6 Dissemination of research 9

1.6.1 Journals 9

1.6.2 Conferences 10

1.6.3 Presentations 10

2 Background and related work 11

2.1 Digital image processing 11

2.1.1 Image formation and representation 12

2.1.2 General applications 14

2.1.3 Image segmentation 15

2.1.4 Object recognition 18

2.1.5 Image classification 21

2.1.6 The discrete cosine transform 24

2.2 System identification 25

2.2.1 Nonlinear system identification models 27

2.3 Detection of the model structure 51

2.3.1 The FROLS algorithm 52

2.4 Computer aided diagnosis 56

2.4.1 CAD for breast cancer 57

2.4.2 Diagnosis performance metrics 58

2.5 System identification models and ANN into CAD for breast cancer . 59

2.5.1 Parametric model-based CAD systems 59

2.5.2 ANN-based CAD systems 62

2.6 Chapter remarks 64

3 Image classification using a 2D-NARX model 67

3.1 Introduction 68

3.2 The 2D-NARX methodology 69

3.2.1 Digital mammogram partitioning 71

3.2.2 Two-dimensional image rendering and representation 73

3.2.3 The NARX model 75

3.2.4 FROLS model structure detection 80

3.2.5 Extraction of feature values 82

3.2.6 Classification and detection 85

3.3 Experiments and results 87

3.3.1 Case of study 88

3.3.2 Setting up of the model parameters 89

3.3.3 Data labelling and supervised learning 90

3.3.4 Resulting image models 92

3.3.5 Classification performance metrics 100

3.3.6 Classification and detection results 100

3.4 Discussion 103

4 Image classification by MSRBF networks and DCT 107

4.1 Introduction 108

4.2 The MSRBF DCT methodology 112

4.2.1 Discrete-time system structuring 114

4.2.2 Traditional RBF and 2D MSRBF neural networks 116

4.2.3 Model structure detection 123

4.2.4 Feature extraction and the DCT 125

4.2.5 Classification and detection 127

4.3 Experiments and results 128

4.4 Discussion 139

5 Image classification by Multilayer-Fuzzy ELM 141

5.1 Introduction 142

5.2 Preliminaries and definitions 145

5.2.1 ELM and multi-input-multi-output RBFNN 145

5.2.2 MIMO IT2-RBFNN and fuzzy logic 147

5.2.3 Multilayer kernel extreme learning 152

5.3 Multilayer fuzzy extreme learning machine 154

5.4 Experiments and results 158

5.4.1 Classification of handwritten digits 159

5.4.2 Breast cancer classification and detection 160

5.5 Discussion 163

6 Conclusions and final considerations 165

6.1 Summary 165

6.2 Conclusions 168

6.3 Future work 170

Bibliography 172

A Examples of study case mammograms 205

A.1 Benign mammogram mdb005 206

A.2 Normal mammogram mdb009 207

A.3 Malign mammogram mdb028 208

B 2D NARX testing results 209

C MSRBF testing results: Tissue-type ratio 1/4 214

List of Figures

1.1 Knowledge areas related to the presented methodologies. 3

2.1 General flow diagram of image processing methods. 12

2.2 (a) Analogue image, (b) digitisation, (c) quantisation. 13

2.3 a) Von Newmann, neighbourhood b) Moore neighbourhood. 18

2.4 Overview of object recognition processing. 19

2.5 Basic concept of classification (taken from [32]). 21

2.6 Basic model of an input-output system. 26

2.7 Basic components of ANN. 31

2.8 Learning process of ANN. 32

2.9 Types of propagation in ANN. 33

2.10 McCullough-Pitt artificial neuron model. 35

2.11 Structural example of DFN (adapted from [122]). 36

2.12 Convolution mask processing over an image. 38

2.13 Max pooling for a 2×2 neighbourhood. 39

2.14 ELM-AE architecture (taken from [153]). 41

2.15 Basic scheme of an RBF neural network. 44

2.16 MF of average students. 47

2.17 MF of taller students. 47

2.18 Three FOUs for the variable *height* (adapted from [165]). 49

2.19 General strategy for detecting the model structure. 52

2.20 Flowchart of medical image processing and CAD systems. Taken from [32]. 57

3.1 2D-NARX general algorithm chart. 70

3.2 Magnification of microcalcifications in mammogram mdb233 (from [212]). 71

3.3 Multiple-input and single-output system. 73

3.4 MISO system modelling of the image field. 74

3.5 Neighbourhood mask scanning movement during the image data collection. 75

3.6 Flowchart of the NARX mapping and polynomial expansion of the nonlinear function. 81

3.7 FROLS's model selection flowchart. 82

3.8 Stimulating the model's behaviour. 83

3.9 Feature extraction by stimulating the model behaviour. 85

3.10 Flow diagram of the classification and detection design. 87

3.11 Typical medio-lateral oblique view of images mdb005 -benign-, mdb009 -healthy- and mdb028 -malign- (from [212]). 88

3.12 Example of a mammogram with several background artefacts (image 274mdb [212]). 91

3.13 The output y in time period k can be explained by combining regressors $u_1(k-1), u_2(k-1), u_3(k-1)$ and $y(k-1)$ in the proposed model. When in the coordinate system y equals to $y(i,j)$, the lagged variables equal to $u_1(i-1,j-3), u_2(i-1,j-2), u_3(i,j-3)$ and $y(i,j-2)$. 94

3.14 ROI from a malign sample (mammogram mdb005 [212]). 96

3.15 Model fitting to data, benign ROI from mdb005 [212]. 97

3.16 Model fitting to data, healthy subimage from mdb009 [212]. 98

3.17 Model fitting to data, malign ROI from mdb028 [212]. 99

3.18 Microcalcification (a) falling into the benign class thanks to sample (b). Subimage (c) falling into the healthy class thanks to healthy sample (d). Malign tumours (e) and (g) falling into the malign class tanks to samples (f) and (h) respectively (images from [212]). . . . 103

4.1 Predicting the unknown system's behaviour via system identification.109

4.2 ROI splitting for a two-fold characterisation. 112

4.3 MSRBF model approximation flowchart. 113

4.4 The shape of the Gaussian function contained in the RBF kernel. . 117

4.5 Multiple-input single-output architecture of a Gaussian RBFNN before the multiscale expansion to be shown in Figure 4.6 117

4.6 Increase in the number of RBF neurons produced by the multiscale approach regarding the architecture shown earlier in Figure 4.5 . . . 119

4.7 Flowchart of MSRBF-based image processing for feature extraction. 125

4.8 Example of 2D DCT information compression in a ROI. 127

4.9 Role of the MSRBF DCT into a classification-based CAD system. . 127

4.10 Accuracy as a function of the presence of dense mammograms in the test set. 134

4.11 Sensitivity and lesion distinction accuracy as functions of the presence of fatty mammograms in the test. 136

4.12 Specificity and NPV as functions of the presence of glandular mammograms in testing. 137

5.1 Fuzzy logic process implicit in the neural network (ROI from [212]). 147

5.2 Multiple-input-multiple-output Interval Type-2 RBFNN 148

5.3 Singleton fuzzification and interval secondary MF which becomes active if $\mathbf{x}_p = x'_l$ for the ith recipient unit of the network (taken from [245]). 149

5.4 Architecture of a Multilayer Kernel Extreme Learning Machine (from [20]). 153

5.5 Fuzzy Autoencoder and MIMO IT2-RBFNN based on Extreme Learning Machines. 155

5.6 Fuzzy Autoencoder and MIMO IT2-RBFNN based on Extreme Learning Machines (mammogram mdb028, from [212]). 161

A.1 Mammogram with a benign tumour. Film mdb005 from the mini-MIAS database [212]. 206

A.2 Mammogram in healthy clinical condition. Film mdb009 from the mini-MIAS database [212]. 207

A.3 Mammogram with a malign tumour. Film mdb028 from the mini-MIAS database [212]. 208

List of Tables

3.1 Average tumour diameter of the MIAS database [212]. 72

3.2 Database mammogram class distribution [212]. 89

3.3 Model of a benign mammogram (1024 × 1024 px, mdb005 [212]). . 93

3.4 Examples of the equivalence between representation systems for the proposed method. 94

3.5 Model of a healthy mammogram (1024 × 1024 px, mdb009 [212]). . 95

3.6 Model of a malign mammogram (1024 × 1024 px, mdb028 [212]). . 96

3.7 Model of a benign ROI (64 × 64 px, mdb005-217 [212]). 97

3.8 Model of a healthy ROI (64 × 64 px, mdb009-184 [212]). 98

3.9 Model of a malign ROI (64 × 64 px, mdb028-182 [212]). 98

3.10 Partition data and initial 2D-NARX results. 101

3.11 2D-NARX tumour detection results. 102

3.12 Comparison of 2D-NARX with previous parametric methods. 102

4.1 Database breast-type distribution [212]. 129

4.2 Two pairs of fit-to-data curves and ERR values. ROI from [212]. . . 130

4.3 Indices for precision validation for the two pairs of model prediction-to-data curves displayed earlier in Table 4.2. 131

4.4 Six ROI pairs, each aligned vertically, and below them the Euclidean distance between their feature vectors. These vectors are obtained from the image model output's DCT compression. Note that the more the visual difference, the larger the gap. Images from [212]. . . 132

4.5 MSRBF DCT performance results by breast tissue-type ratio. . . . 133

4.6 False (positive and negative) cases during testing. ROIs from [212]. 135

4.7 MSRBF DCT overall performance results. 137

4.8 Comparison of the MSRBF DCT method with previous work. . . . 138

5.1 Comparison between the new ML-FELM and previous ML networks in the MNIST database. 160

5.2 Performance comparison between the proposed ML-FELM and previous machine learning techniques for breast cancer detection. . . . 162

B.1 2D NARX model test results 1-15 209

B.2 2D NARX model test results 16-40 210

B.3 2D NARX model test results 41-64 211

B.4 2D NARX model test results 65-86 212

B.5 2D NARX model test results 87-100 213

C.1 MSRBF: Test tissue-type ratio 1/4, Results 1-9 214

C.2 MSRBF: Test tissue-type ratio 1/4, Results 10-35 215

C.3 MSRBF: Test tissue-type ratio 1/4, Results 36-61 216

C.4 MSRBF: Test tissue-type ratio 1/4, Results 62-87 217

C.5 MSRBF: Test tissue-type ratio 1/4, Results 88-113 218

Nomenclature

ARMA	AutoRegressive Moving Average
AE	AutoEncoder
ANN	Artificial Neural Network
CAD	Computer Aided Diagnosis
DCT	Discrete Cosine Transform
ELM	Extreme Learning Machine
ERR	Error Reduction Ratio
ESR	Error to Signal Ratio
FAE	Fuzzy Auto Encoder
FLS	Fuzzy Logic System
FROLS	Forward Regression Orthogonal Least Squares
FS	Fuzzy System
IT2	Interval Type-2
MF	Membership Function

MIMO	Multiple Input Multiple Output
MISO	Multiple Input Single Output
ML-FELM	MultiLayer-Fuzzy Extreme Learning Machine
ML-KELM	MultiLayer-Kernel Extreme Learning Machine
MSRBF	MultiScale Radial Basis Function
NARMAX	Nonlinear AutoRegressive Moving Average with eXogenous Inputs
NARX	Nonlinear AutoRegressive with eXogenous Inputs
OCI-ELM	Orthogonal Convex Incremental-Extreme Learning Machine
RBF	Radial Basis Function
ROI	Region of Interest
SVM	Support Vector Machines

Chapter 1

Introduction

1.1 Background

The visual sense is the primary way that humans experience and get in touch with reality [1]. In humans, the visual cortex is the centre for the processing of visual information since it extracts the necessary content to perform spatially complex tasks [2]. Analogously in artificial intelligence, image processing retrieves the characteristic information contained in digital images through algorithms mainly based on statistics and numerical analysis. Image processing aims at (a) the strengthening of the representation quality, or (b) extracting useful features from the representation to complete learning-related goals such as detection and classification [3]. However, unlike the visual cortex, image processing is capable of obtaining both high and low-level features, which correspond to human and machine comprehension respectively. Examples of relevant application areas are security and defence, remote sensing, microscopy, robotics and medicine [4].

Given the growing availability of the volume of digital information derived from recent advances in information technology, the improvement in the capacity to recognise high-quality features from visual data is a central problem in image processing. However, this is often challenging because of inherent problem diffi-

culties. The authors in [4] found that manoeuvres on digital images are not linear, although linear procedures can approximate these systems in mild circumstances. Also, linear manipulations in digital images may lead to poor results when there is noise with different statistics to Gaussianity [5]. Similarly, the authors in [6] found that linear filters in image processing can miss important image features, such as borders separating objects from the background.

This problem has been approached from different perspectives, such as artificial neural networks [7], [8], [9], and linear system identification models [10], [11], [12]. However, before the work presented in this thesis, the system identification approach for feature extraction had only be incorporated by linear models, despite the nonlinear ones have proven excellent results in the approximation of several dynamic problems [13], [14], [15], [16], [17]. This work makes this incorporation and also presents a new neural network design based on autoencoders, which are known by their bottleneck architecture designed to retrieve feature values efficiently [18], [19], [20]. The new network combines the autoencoders with fuzzy logic [21], which adds robustness to the system in the presence of uncertainties, a handy advantage in classification problems. Figure 1.1 portrays the interrelations between the new methods accordingly to their knowledge area.

The new methods take the form of Computer Aaided Diagnosis (CAD) frameworks for the relevant problem of breast cancer detection. According to [22], the rates of female deaths for breast cancer in the world are still in a terrible situation in spite these have decreased in developed countries thanks to new detection technologies. However, in the United States, breast cancer is placed at the first place of new cases by cancer type and lays as the second cause of cancer deaths in women by 2017 [23].

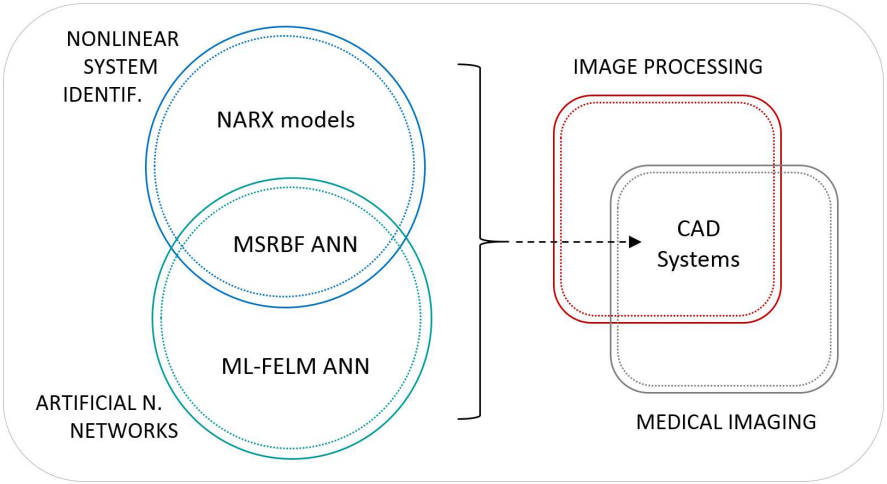


Figure 1.1: Knowledge areas related to the presented methodologies.

The global aim of the thesis is to make available new and more flexible image processing methods able to deal with corrupted spacial features by taking advantage of forefront models based on nonlinear analysis, while it presents the breast cancer detection problem as a study case to validate the methodologies.

1.2 MOTIVATION

Information technologies are evolving towards increasingly more efficient methods, both in quality, as in processing volume and speed. The processing of digital images is one of the branches that has received the most of attention, not only because of its high presence in the media but also because of the growing role it plays in decision-making, among others, in national security, surveillance, geographic systems, microbiology and medicine [3],[24].

One of the primary objectives of image processing in this kind of applications is the extraction of high and low-level features. High-level features are recognis-

able by the human eye. The detection of the low-level features takes place at the algorithmic level. The advance of the capacity of image feature extraction technologies in recognising visible and non-visible elements and/or reducing the gap between levels is, therefore, of increasing importance [3].

This thesis contributes to solving this problem via three new classification methods based on nonlinear system identification models and artificial neural networks for the extraction of high and low-level features since the proposed techniques lay upon extraordinarily flexible approximation function methods such as the NARX polynomial model and efficient structured neural networks based on Gaussian functions.

Among the real-life disciplines mentioned earlier, the medical field is one that has benefited the most from image feature extraction, specifically through CAD technologies [25]. In this work, the proposed methods are coupled to classifiers to integrate CAD systems as a *second opinion* tool to attack the breast cancer problem, which is especially relevant not only because of the high mortality rate linked to it but also by the positive healing potential when it is detected in early stages [26].

1.3 Aims and objectives

1.3.1 General objective

To present new digital image feature extraction and breast cancer detection frameworks to increase the availability and scope of existing approaches by leveraging and combining the advantages of nonlinear system identification, simple and multilayer artificial neural networks.

1.3.2 Specific aims

- To use nonlinear system identification models in the form of the Nonlinear AutoRegressive with eXogenous inputs (NARX) model and the Generalized Multiscales Radial Basis Function (MSRBF) networks to build up models capable of capturing the two-dimensional elements contained in images.
- To design a method capable of leveraging the image models (built through system identification) to extract representative feature values.
- To transform into a reduced amount of coefficients the most valuable information extracted from the image models to lessen the computational burden in the classification process.
- To combine the new digital image feature extraction procedures with a suitable classification algorithm to build up new CAD systems for breast cancer.
- To take advantage of fuzzy logic systems, stacked autoencoders and radial basis function networks to integrate a new CAD system for breast cancer.

1.4 Thesis overview

The thesis content is structured as below:

- Chapter 2 reports a review of the related work concerning the frameworks proposed in the following chapters. It goes over digital image processing, system identification and computer-aided diagnosis (CAD) with an stress in the background theory of parametric models and neural networks within system identification. There is also a theoretical emphasis on the NARX

models, radial basis function, multilayer neural networks and fuzzy logic systems since these are the direct antecedents of the proposed methods.

- Chapter 3 presents the polynomial nonlinear autoregressive with exogenous inputs (NARX) model as a nonlinear system identification model for image feature extraction. The framework aims to seize the NARX capability to portray dynamic systems into models, so complex structures within images can also be adequately retrieved. The chapter reports as well a polynomial NARX formulation for digital images, termed 2D NARX. Its solution takes place through the forward regression orthogonal least squares (FROLS) algorithm and the *k-means++* clustering method. Also, the polynomial NARX model takes shape as a CAD system for breast cancer detection. Experiments show the capacity of NARX-FROLS to derive image models and the effectiveness to classify mammograms compared to previous CAD methods based on system identification.

- Chapter 4 presents the multiscales radial basis function network (MSRBF) in digital image processing. The MSRBF network combines a single hidden layer structure which is highly competent to describe complex systems since its efficiency outstands in the identification of real-life dynamical systems. The objective is to produce concrete image models thanks to the inclusion of scales in the hidden neurons, while the forward regression orthogonal least squares (FROLS) algorithm selects the model structure. The discrete cosine transform (DCT) converts the model output into highly compacted feature values. A mathematical modelling was done to adapt the MSRBF network as an image processing method by viewing the image as an input-output system. To evaluate the method the problem of breast cancer detection in X-ray mammography was adopted. Classification results show that the new

characterisation method helped reach a very competitive diagnostic accuracy among other measures. The MSRBF network could also generate highly accurate images models.

- Chapter 5 presents a Multilayer Fuzzy Extreme Learning Machine (ML-FELM) that is based on the functional equivalent between the Radial Basis Functions (RBF) and Fuzzy Logic Systems to classify the mammograms. The ML-FELM is a fast forward multilayer neural structure whose parameter identification consists of two main phases. In the first one, Fuzzy Autoencoders (FAEs) intervene for the extraction of high-level image features. In the second phase, a fuzzy RBF is implemented for the classification of the features extracted by the FAEs. The use of some other automated ELM methodologies served to evaluate the performance of the proposed ML-FELM. Results of the proposed ML-FELM applied on the MNIST, and mini-MIAS data sets show a significant trade-off between accuracy and model simplicity.

- Chapter 6 summarises this thesis and reports the conclusions and directions for future research.

1.5 Main contributions

This thesis explores new frameworks for image processing concerning feature extraction and classification to enhance CAD systems based on a nonlinear analysis. The most significant research contributions are described as follows:

1. **Presentation of the polynomial NARX model in image feature extraction**. The NARX performs as a flexible-order system identification model for image feature extraction for the first time. The idea is to take

advantage of the proven capability of the polynomial NARX representation to capture the subtle elements of nonlinear dynamic systems, so both smooth and corrupting spacial features within images do not be omitted or shortly described. The method includes the adaptation of the 2D image format to an input-output dynamic system representation and the model's stimulus-response design for feature extraction. The model structure detection materialises via the FROLS algorithm.

2. **A polynomial NARX-based CAD system**. The polynomial NARX model for image feature extraction appears for the first time in a CAD framework for breast cancer detection. The *k-means++* algorithm acts as a clustering method that links the training and the testing vectors to produce a pre-diagnosis to be monitored by medical evaluation. Experiments show the capacity of the system to derive image models and the effectiveness to classify mammograms compared to previous CAD methods based on system identification models.

3. **Presentation of the MSRBF network in image feature extraction**. The direct use of the MSRBF networks within image processing takes place in this research for the first time. This network holds an efficient and straightforward structure initially designed for the identification of input-output systems. The aim is to use it to get concise and accurate image models thanks to the flexibility provided by the inclusion of scales in the Gaussian functions. The FROLS algorithm solves the model structure detection problem. After the model building, the discrete cosine transform (DCT) compresses the energy of its output into a few coefficients to form feature vectors of high quality.

4. **A MSRBF-based CAD system**. The MSRBF network addresses the problem of breast cancer detection in X-ray mammograms for the first time. It works along with the DCT transform to take advantage of their joint capacity for flexible model approximation and feature extraction. Classification results show that the new characterisation method helped reach a very competitive diagnostic accuracy, sensitivity, specificity, positive predictive value and negative predictive value.

5. **Presentation of the ML-FELM neural network**. The Multilayer Fuzzy Extreme Learning Machine (ML-FELM) bases its power on autoencoders neural networks, radial basis function networks (RBFNN) and fuzzy logic systems. It aims to classify digital images in general, and regions of interest from mammograms as a CAD system. The ML-FELM uses autoencoders for the extraction of image features and Fuzzy-RBFNN for the classification of the encoded features. Results on data for handwritten digits and breast cancer detection show a high model accuracy, while several other methodologies are used to compare its performance.

1.6 Dissemination of research

1.6.1 Journals

The listed research was reported to the following journal:

- C. Beltran Perez, A. Rubio Solis, H.-L. Wei. "A Multilayer Fuzzy Extreme Learning Machine for Breast Cancer Image Classification". Pattern Recognition Letters, Elsevier, (2018). Manuscript submitted for publication.

1.6.2 Conferences

The listed research was reported and presented in the following conferences:

- Carlos Beltran Perez and Hua-Liang Wei. "Digital Image Classification and Detection Using a 2D-NARX model". In 2017 23rd International Conference on Automation and Computing (ICAC): Addressing Global Challenges through Automation and Computing, IEEE, 2017.
- Carlos Beltran Perez and Hua-Liang Wei. "Image Classification Using Generalized Multiscale RBF Networks and Discrete Cosine Transform. In 2018 24th International Conference on Automation and Computing (ICAC): Improving Productivity through Automation and Computing, IEEE, 2018.

1.6.3 Presentations

The presentation of the listed research took place in the following symposiums:

- Beltran Perez, C. "Enhanced Computer Aided Diagnosis for Breast Cancer Imaging Based on System Identification Procedures". In: The University of Sheffield, Engineering Research Symposium, 2018.
- Beltran Perez, C. "Digital Image Classification by Using a 2D-NARX Model". In: The University of Sheffield, Sheffield Neuroscience Conference 2017.

Chapter 2

Background and related work

This chapter presents a review of the work related to the contributions of the thesis. It makes an overview of digital image processing, system identification and computer-aided diagnosis. It highlights the background theory of nonlinear parametric models, neural networks and fuzzy systems in the context of system identification and image processing.

2.1 Digital image processing

Digital image processing is an interdisciplinary and ubiquitous branch of signal processing and computer science that uses computer algorithms to enhance specific features or extract relevant information of digital images. This area encompasses a plethora of tasks that go from low-level processing, for instance, contrast enhancement, medium-level processing such as edge detection and thresholding, to high-level tasks involving complex algorithms based on statistics or system identification to produce concise and efficient image descriptions [27],[28].

As regards the processing order of an image taken from a real-life object, the hierarchy of tasks goes from the capture, digitisation, quantisation, prepro-

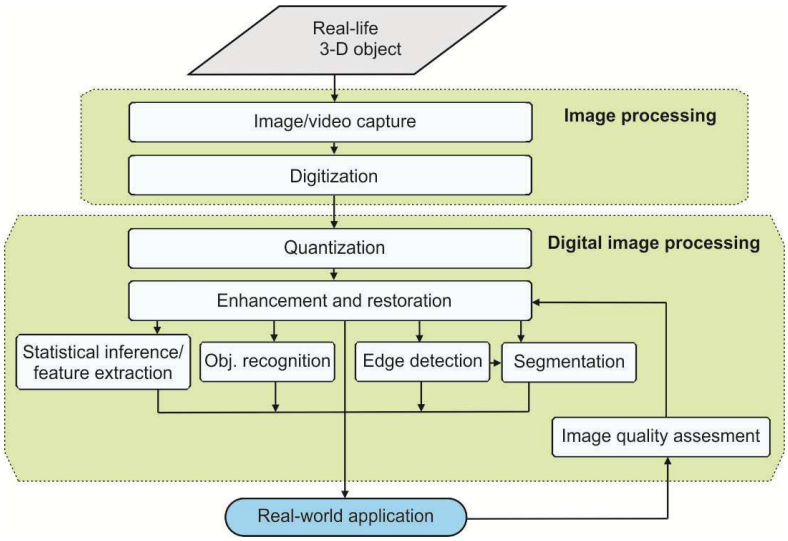


Figure 2.1: General flow diagram of image processing methods.

cessing and medium to high-level processing (Figure 2.1) [5].

Digital image processing is especially important to society because it is highly interdisciplinary and can be a tool to solve problems from different areas such as medicine, astronomy, engineering and criminology and at the same time it uses concepts from other disciplines as optics, radiometry, geometry and computer sciences [27],[1].

2.1.1 Image formation and representation

The sense of vision is perhaps the most crucial perception system of human being, as it allows them to quickly receive information from the environment and use it as a basis for fundamental and complex tasks like moving, balance and protect themselves from external threats. In the same way, the study of visual observation and its applications has been a significant part of science since its inception. Nonetheless, the scope of this discipline has grown thanks to two historical contributions:

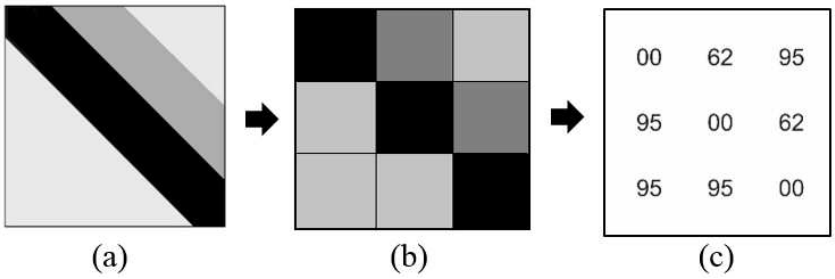


Figure 2.2: (a) Analogue image, (b) digitisation, (c) quantisation.

the photography by Louis Daguerre in 1839 [29], and the digital camera, by Steven Sasson in 1975 [30].

The invention of Daguerre, known today as analogue photography, was able to capture images using a lens within an opening of a light-tight box with a film with silver halides in the opposite end. This advance allowed for the first time to document an image objectively beyond technical drawings, and written descriptions [27]. Meanwhile, the digital camera of Sasson used the lens and light-tight box used by Daguerre, but electronic photosensitive sensors that collectively captured the optical wave fields as a continuous function $f(x, y)$. Thereinafter, the captured data was sampled and organised into a two-dimensional grid easy to be quantised (in other words, converted into numerical values) based on a sampling-based feature termed image intensity (Figure 2.2).

Since then until today, these discoveries have enabled computer science and computer vision, among other disciplines, to adopt and develop fast algorithms for many tasks such as evaluating, exploring, classifying and recognising, based on mathematical calculations and statistical processes performed on the approximated images [4].

The intervention of computer science accompanied by the development of

increasingly fast computer hardware, higher-resolution acquisition techniques and superior storage capacity during the last four decades have transformed the way today's scientist approximate visual analysis, that mainly comprises three strongly interrelated areas: digital image processing, machine vision and image analysis [4],[27].

Nowadays digital image processing encompasses the algorithms that process higher-dimension signals (2-D and 3-D) to complete specific tasks. Machine vision is related to image acquisition techniques that obtain information from both visual and non-visual wave fields. Image analysis comprises algorithms that automatically measure, examine and describe two and three-dimensional attributes and quantitative features from images [31].

2.1.2 General applications

Image processing applications contribute in general to storing, refining and evaluating visual information, in motion or captured, from real problems such as microscopy, iris recognition or object detection to facilitate or automate human decision-making processes [1].

To exemplify the above, comprehensive overviews of digital image processing techniques are for medicine [32] and [33], where noise reduction, object detection, image segmentation and feature extraction play a common task. In CAD systems the work in [34] and [35] stand out as relevant compilations. In microscopy [36], where a consistent interpretation of micrographs takes place. In astronomy [37], where high spectral and spatial resolutions are central requirements. In geography [38], where geographical information (GIS) and remote sensing systems play a central role.

Regarding real-world applications and applied sciences, digital image processing has tackled a number of diverse everyday problems in nearly all areas. Prominent practical applications of digital image processing go from surveillance [39],[40],[41], where video analysis and object tracking are mainly used to prevent and combat delinquency, remote sensing [42],[43],[44], where the primary objective is to classify map zones by visual identification, plant identification [45][46], by computing digital images of plants to construct a classification framework, robot guidance [47],[48],[49], where the visual environment is used as a reference to reduce trajectory deviations, and flow visualization [50][51],[52], where image analysis of flows delivers quantitative information that is useful as a measuring procedure or feedback. Numerous additional applications of digital image processing can be found in [3],[4],[53],[24],[1].

2.1.3 Image segmentation

Image segmentation is an essential step in image processing that divides the image into two or more distinctive and homogeneous regions, each containing strongly linked pixels that cooperatively represent a real-world entity [4],[24]. It is essential because various image analysis and feature extraction processes depend on segmentation to intensify the objects, patterns or features to be extracted and matched. The section shows the most common and attractive image segmentation techniques in the literature.

Image thresholding segmentation usually aims to determine, as robust as possible, a difference between the object or objects and the background. One of the most relevant and multi-cited thresholding algorithms is the clustering-based Otsu method [54] which finds the mean of the average levels for two different classes

to select the threshold value. At first, the histogram of intensities is computed and normalised.

From the histogram, every grey level in the image is used as possible threshold value t to compute the variances of the foreground (intensity levels above t) and the background (intensity levels equal or below t) for each candidate threshold. Finally, it is selected the threshold with the minimum sum of upper and lower variances. Though the method aims at optimal thresholding, it tends to establish the value in function of the entity with larger variance within the class.

Between the region-based segmentation tactics, the unsupervised region growing algorithm is an attractive bottom-up alternative that, in spite of its proven efficiency, only a small number of researchers have employed it [55]. In general, the algorithm takes seeds as first regions which grow iteratively if the adjacent pixels are unclassified. This annexing process is also carried out between regions by a tagging strategy that compares them to investigate if these are similar, in which case the regions merge each other.

Exciting and recent work on image retrieval segmentation is in [56] for land-type recognition in satellite images, showing convincing results in grayscale and colour images. In conclusion, though the image retrieval algorithm is exceptionally efficient in noisy images, it is not as exploited in the literature as it could. A drawback of the method is that it relied on the accuracy of the seeding process.

In [57] a hybrid colour image segmentation method is adopted with region growing algorithm, cloud model and seed selection via Harris corner detector. The latter is a probabilistic method low-sensitive to rotation, noise and brightness variations. Also, the cloud model is adopted to automatically grow the seed-centred regions adopting or rejecting adjacent pixels in correspondence to a threshold es-

tablished by the cloud. This threshold takes into account the expected value, entropy and hyper entropy to compare the values between regions. The method resulted fast and precise but presented some over segmented regions.

The graph cut algorithm, a combinatorial optimisation algorithm was used in image segmentation for the first time in [58], by modelling the array of pixels of the image as an undirected graph where the adjacencies between pixels are the edges and the pixels as nodes. The method employs seeds, previously established by the user, as hard constraints and cost functions between pixels as soft constraints to choose cuts between those pixels with the lowest reciprocal cost.

The first author of [58] further presented the graph cut approach [59] for medical image segmentation for cardiac magnetic resonance data. In this version, directed and undirected graphs take place along with additional constraints in the optimisation model to correct over segmented zones. The graph cut segmentation in medical images presented efficiency, robustness and realistic modelling of specific requirements. The method was presented depending on the user input to establish seeds.

The work in [60] presents a newer approach to image segmentation using cellular learning automata (CLA) as a skin detector. CLA is a discrete space-time system composed of identical squared points which state changes as a function of a simple rule (usually reward-punishment) which in turn depends on the cell environment. All cells have the same environment size, which typically follows one of the morphological configurations depicted in Figure 2.3 (representation taken from [61]).

CLA is recent to image processing, being edge detection its most common application. In the mentioned skin segmentation approach, the CLA rule decides

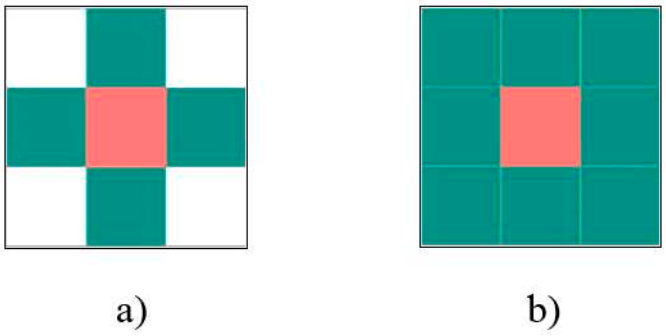


Figure 2.3: a) Von Newmann, neighbourhood b) Moore neighbourhood.

whether a neighbourhood is a skin or not by using a probability map. The probability map propagates in all directions until a decision for an entire region takes place. After some iterations, the procedure stops when the whole system converges. Experimental results showed good performance of the algorithm compared to previous skin detectors regarding false positives. However, the texture data of the skin is not included while better performance is still desirable in low contrast images.

2.1.4 Object recognition

Object recognition processes aim at finding entities within the images taking into account pre-specified patterns in a supervised or unsupervised way [62].

The recognition process converges from two parallel lines: learning and classification. The general chart of this process, taken from [58], is in Figure 2.4. As reported in the next section, this family of methods traditionally depend on statistical methods to classify and compare from previous and new information, so that the most common pattern recognition approaches in the literature are statistical-based. The most representative of them are below.

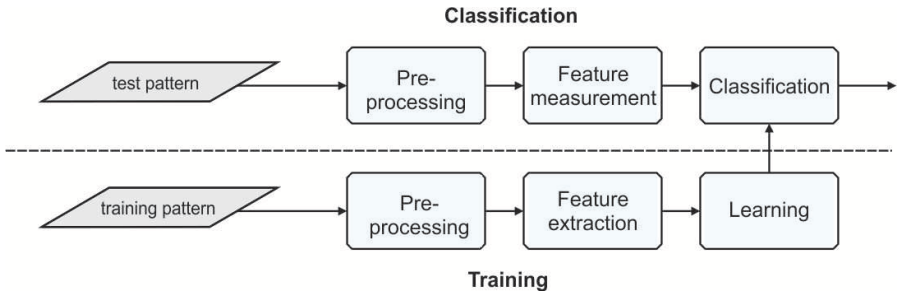


Figure 2.4: Overview of object recognition processing.

Template Matching is a statistical approach aimed at finding a prototype or template of a point, saliency or shape (learning process from a training set) and store it to compare later its similarity with new unknown entities (matching and classification process). This task takes place by optimising the correlation (a similarity index) during the learning process. Applications of template matching include medicine, remote sensing and three-dimensional recognition.

A relevant application of template matching in X-ray computed tomography is [63] where an elastic deformation of the regional variant data occurs during three coarse-to-fine stages that fix the more significant disparities at each step with the aim of improving the local similarity. The iterative process results in models or templates with an increased resolution that mach the original data. Experiments in X-ray brain scans show that elastic matching effectively detected the position and shape of three different brains from 3 viewpoints. A disadvantage of this work is the high computational demand of the method.

The statistical object or pattern recognition approach [64] represents the objects in the image as a set of measurements or features that are separated by boundaries in the measurement space to define different classes. Probability distributions serve as standards to compute the boundaries, and their training or computation is a priori. In the first case, an automatic training stage based on

classification can be introduced as a criterion, for instance, MSE. Another way to determine boundaries rests on suggesting several boundaries, analyse them and discriminate them by taking into account the classification of patterns. As mentioned, this recognition methodology needs as part of the inputs probability distributions which must be specified by the user or inferred by additional methods.

Syntactic matching is a robust recognition approach designed to identify complex entities by adopting a hierarchical decomposition of the more complex patterns into more straightforward and more uncomplicated subpatterns until getting to the *primitives*, the basic building blocks. In this way, numerous patterns of high complexity end up being defined by a few primitives and a set of rules.

The syntactic matching approach arises in image processing in the identification of written characters, as in [65] where grayscale frontal photographs of cars are processed as inputs to recognise the plate number code. The region growing algorithm, referred before in this chapter, is used in the first two stages to perform a segmentation which helps to isolate the car plate from the remaining image area in the first place, such that the code number left black, the plate background white and the rest of the image black. Then a second region growing segmentation is processed to isolate the code number in black within a background in white. The pattern recognition strategy then focuses on finding the optimal set of primitives by considering string conflicts, such as Y and T in the form of possible representations for each letter and number. The tests demonstrated an accuracy of 95%. However, two main drawbacks remain to be solved: a problem in the segmentation of noisy objects and the possible combinatorial explosion when the number of primitives is too large.

2.1.5 Image classification

Image classification is very active in science and engineering because its development enables computers to support humans visual-related decisions [62]. It unifies machine learning and image processing as it involves the statistical learning process of a desired output or pattern in digital images. This kind of procedures are intended to (a) extract information from an image set to generate classes or subsets of images with similar features or (b) extract information classes within a single image. Depending on the application context and the human-machine interaction degree during learning, the image classification process is supervised, semi-supervised or unsupervised. However, an inherent difficulty is common to this problems: the fact that in numerous cases the image data are not linked to the classes of interest, in which case it is required to conduct a careful and comprehensive practice [27].

Figure 2.5 shows a flow diagram of the conventional classification process, where the inputs may vary in format, but the output is always a discrete value that denotes the class. Image classification is essential in many areas nowadays. Relevant examples of these areas include:

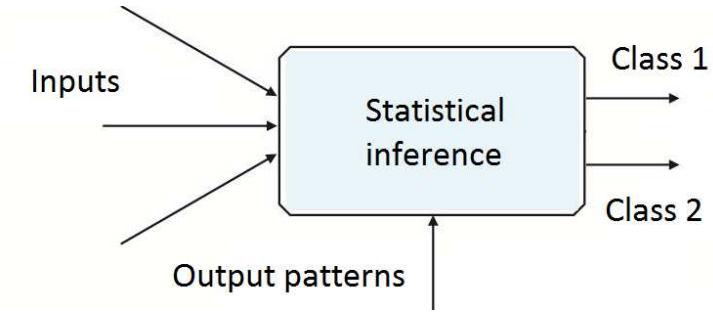


Figure 2.5: Basic concept of classification (taken from [32]).

- Remote sensing, which uses multispectral or multimodal satellite images for studies such as urban planning and meteorological control [66].
- Automatic visual inspection, which focuses on improving industrial processes such as defective component control [67].
- Military surveillance, for the location of strategic objectives [68].
- Tracking of moving objects, which can be recognition-based or motion-based [69].
- Image compression for storage reduction and transmission [70].
- Biomedical imaging, which is used mainly for the diagnosis of medical images such as heart, lung and breast [71],[72].

Due to the research objectives, the following section highlights the classification of biomedical images and computer-aided diagnosis.

Image classification in biomedical imaging

The primary target of classification In biomedical imaging is to associate patterns of measurements with a specific disorder or condition, or in other words, determining a disease from a collection of registers. The most usual cases are the simultaneous monitoring of the patients' situation and when the available information is very extensive or very complex for easy understanding by the specialist [32]. Below are the most representative classification techniques of biomedical image classification along with related work.

- *Linear Discriminators.* The Fischer discriminant analysis [73] also known as linear discriminators, is one of the simplest classification methods, since

it resolves by dividing the data set with a line for two dimensions, a plane for three dimensions and so on. Given that the method traces one single boundary, two possible results are always available. A weighted sum is made to know the class from which any set of observations belong. The weights are estimated a priori by training, and the inference takes place by comparing the weighted sum of the set to a threshold value.

A typical instance of this approach in CAD is that of [74] in which ultrasonic image analysis is used to diagnose hepatic fibrosis. In this case, a set of textural parameters result from a set of healthy and fibrosis-related images. Then textural data are captured from new images to feed the linear discriminator. The method attained an overall diagnostic accuracy of 75%.

- *Support Vector Machines.* The work of [75] introduced the support vector machine (SVM) as a machine learning method for classification and prediction aimed at avoiding the data overfitting while maximising the prediction accuracy. The structural risk minimisation concept is used to improve the generalisation of the model (or solve the overfitting problem) by equilibrating the model fitting versus the model complexity. SVM uses distances or margins between some class data points and sets of constructed subspaces (hyperplanes) to attain a useful separation [76].

  In [77] SVM is used in CAD to classify solid breast tumours after a previous segmentation step. The SVM used 5 different datasets for training and six morphologic features for classification criteria: form factor, roundness, aspect ratio, convexity, solidity and extent. The performance of SVM in almost all morphologic features surpassed 90%. However, the recommended approach in breast images depends on texture analysis, so its use is restricted to ultrasound images since these are rich in textures.

- *K-means and k-prototypes.* Within cluster analysis, the k-means clustering classifier [78] involves a different training tactic by taking representative data as *prototypes* or data centres. This way, the number of k-centres is equal to the number of selected prototypes. After the user performs this selection, there is a data distribution according to the closest Euclidean distance to centres.

  The k-means algorithm has several extensions, including the learning vector quantisation method [32] to find an improved centre positioning. The k-prototypes algorithm [78] integrates the clustering of both numerical and categorical data. Two CAD applications of the k-means algorithm in breast cancer diagnosis are [10] and [79]. In both works, the classifier is used after the ARMA and ARIMA parametric feature extraction. The first case reported a diagnosis accuracy of 93.8% and the second case a 95%.

2.1.6 The discrete cosine transform

As stated in Chapter 1, one of the particular objectives of this work is to compress the digital image information to lessen the computational burden during classification. Image compression is also known as *coding*. Image coding generally involves a transformation function, being the discrete cosine transform (DCT) the most used thanks to its advantages of high energy compression in very few coefficients [24].

The discrete cosine transform [80] is an algorithm that represents images using integer coefficients obtained through a discrete type transformation based on cosine functions. The removal of the correlated coefficients takes place by employing the Karhunen-Loeve transform, involving an orthogonalisation step to generate a series of uncorrelated values. Finally, a sorting of values proceeds in

descending order to concentrate the energy. The DCT compression is regarded as *lossy* because the least essential frequency fragments left discarded in favour of the first k most relevant values, which are later used to reconstruct the picture.

The DCT positions itself as a standard coding method. The most popular standard mode for digital image compression is the Joint Photographic Experts Group (JPEG), where the primary coding system is the DCT [70],[24]. In real-time coding, the DCT acts in video streaming and streaming services, such as the H.264 / MPEG-4 Advanced Video Coding (AVC) encoding used by YouTube [81],[82].

Other uses of the DCT in image processing are object recognition and classification, including biometrics applications such as face recognition. This circle of methods utilise the DCT compression in two common ways: as a preprocessing step to reduce the input image dimensionality of a machine learning classifier [83],[84] or as a feature extraction algorithm that feeds a subsequent distant-based classifier [85],[86],[87].

2.2 System identification

The identification and modelling of complex systems has an essential role in science and engineering because it fills the gap in cases where we know the input-output behaviour of a system but we do not have a mathematical model to understand and predict its changes in future. System identification techniques regard previous observations of a system as explanatory variables to be processed to obtain a trustable mathematical representation [88]. A standard representation in system identification depicts dynamic systems as input-output models. Figure 2.6 shows a basic scheme of a single input and single output (SISO) unknown system.

System identification involves a plethora of techniques that depend on the nature of the problem faced. Examples of relevant applications include Stock prices [89],[90], weather prediction [91],[92], speech recognition [93],[94], pattern classification [95],[96], and aircraft dynamics [97],[98].

The standard processing flow of system identification methods for the recognition of a new model comprises, explicitly or implicitly, the following modules in progressive order:

(A) The choice of the model to be used, which is the dynamic representation of the system.

(B) The corresponding structuring of the available data (system inputs and outputs) in the form of a linear or nonlinear function, according to the selected model.

(C) If unknown, the selection of the model structure.

(D) Model approximation or parameter estimation in the case of parametric models.

(E) Model validation.

Concerning the classification of the available system identification models, it is common to consider linear and nonlinear and parametric, nonparametric and

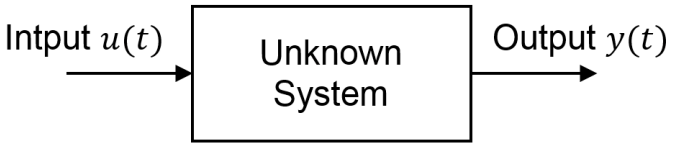


Figure 2.6: Basic model of an input-output system.

hybrid models. In this regard, nonlinear models are nowadays more popular since their increased flexibility help to describe a broader range of physical systems compared to their linear counterpart [99]. Parametric expansions such as NARX models [100], rational NARMAX [101], Volterra series models [102] and block structured models [103] mostly use a discrete-time dynamic understanding of systems as it allows to identify transparent models and the explanatory variables which most minimise the fit-to-data error.

When the goal is merely to approximate the system and the model transparency can be sacrificed, nonparametric system identification methods such as radial basis functions [104], multi-layer neural networks [105], wavelet functions [106], fuzzy logic [107] or nature-inspired metaheuristics [108] can be convenient choices [109],[88].

2.2.1 Nonlinear system identification models

Models are a fundamental part of systems engineering. Their correct identification allows to analyse, understand and predict real systems as well as develop new solutions or even prevent future threats. This section makes a review of system identification models with an emphasis on the techniques which served as a basis for the image processing methods proposed in this work: the NARX model, radial basis function networks, artificial neural networks and fuzzy logic systems.

The NARX models

The Nonlinear AutoRegressive with eXogenous inputs (NARX) model is a special case of the more general class of NARMAX models, an input-output discrete-time representation for a great variety of linear and nonlinear systems proposed in

[100] and [110] designed to identify complex systems. The NARX model approach has proven to be highly accurate to identify a wide range of nonlinear problems [88],[15],[16],[17]. NARX uses a general assumption: the output $y(k)$ is explained by an unknown nonlinear function$F[\cdot]$ depending on sequences of lagged system inputs and outputs to be identified and by a noise sequence $e(k)$:

$$y(k) = F[\cdot] + e(k) \tag{2.1}$$

In the NARX model case, the general representation is:

$$y(k) = F[y(k-1), ..., y(k-n_y), u(k-1), ..., u(k-n_u)] + e(k) \tag{2.2}$$

in which the function $F[\cdot]$ can be linear or nonlinear, $y(k)$ and $u(k)$ the the sequences for the system input and output and $e(k)$ an independent, additive noise sequence.

The expansion of the NARX function can be polynomial, rational or based on neural networks, radial basis functions, wavelet models, fuzzy sets, among others [88]. Of these, the polynomial class is the most popular given its decomposition capability, transparency and readability, qualities that match efficiently to robust structure detection and parameter detection algorithms, offering on the way in-

terpretability and the qualitative performance of the input and output variables [111]. The polynomial expansion of the NARX model is:

$$\begin{aligned} y(k) = \theta_0 + \sum_{i_1=1}^{n} \theta_{i_1} \mathrm{x}_{i_1}(k) + \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \theta_{i_1 i_2} \mathrm{x}_{i_1}(k) \mathrm{x}_{i_2}(k) + ... \\ \sum_{i_1=1}^{n} ... \sum_{i_\ell=i_{\ell-1}}^{n} \theta_{i_1 i_2 ... i_\ell} \mathrm{x}_{i_1}(k) \mathrm{x}_{i_2}(k) ... \mathrm{x}_{i_\ell}(k) + e(k) \end{aligned} \tag{2.3}$$

where k is a discrete time unit ($k = 1, 2, \cdots$), ℓ is the maximum nonlinear degree, $\theta_{i_1, i_2, ..., i_m}$ are the model parameters and $n = n_y + n_u$, being n_y and n_u are the maximum lags for the system output and input. The vector of basic regressors is as follows:

$$x_m(k) = \begin{cases} y(k-m) & 1 \leq m \leq n_y \\ u(k-m+n_y) & n_y + 1 \leq m \leq n \end{cases} \tag{2.4}$$

with $1 \leq m \leq \ell$. Notice that the estimation of the model parameters $\theta_{i_1, i_2, ..., i_m}$ is normally carried out separately after the structure detection. Besides, the exhaustive combination of all regressors in $x_m(k)$ takes place during the polynomial expansion in (2.3) from degree 1 up to the maximum non-linear degree ℓ to shape the candidate model terms. The total number of terms M in the pool of candidates is:

$$M = \frac{(n+\ell)!}{n!\ell!} \tag{2.5}$$

in which n_y and n_u are the maximum lagged observations for the input and the output, and $n = n_y + n_u$. From this viewpoint, each term of the expanded poly-

nomial can be seen as a candidate term to be included in a final model utilising a model structure detection algorithm (Section 2.2.3).

The use of polynomial NARX models have yielded multiple examples in the understanding of diverse real-life complex systems [88],[111],[112],[17]. More applications include the modelling, understanding and forecasting of atmospheric dynamics, as in [13], that reports a study seeking to solve a near-Earth magnetic disturbance prediction problem. The authors bring together evidence revealing that Earth's northern hemisphere temperature is a function of the solar wind speed and the dynamic solar wind pressure, among other variables. To produce mathematical models capable of finding the most influential independent variables, the NARX model was feed with data collected at international geoscience agencies to predict future instances and prevent future threats.

Another precedent is [14], where the meridional overturning circulation of the Atlantic Ocean is effectively modelled via the NARX model to detect the causes of the reduction of its strength in recent years. During the modelling, atmospheric and oceanic density were listed as model inputs while the circulation strength was considered the output. Results suggest that the circulation reduction is due to both a seasonal variability of the sea current strength and different density between the northern and southern hemisphere sea water.

As regards hybrid models, in [113] the ATM cash demand is predicted by the NARX and the NARMAX models, each combined with ANN and support vector machines (SVM). Results showed that the NARX model coupled with ANN obtained the better demand predictions, while none of the NARMAX models produced substantial improvement despite their higher complexity in being solved.

Artificial neural networks

Artificial neural networks (ANN) emerged as a dominant class of models whose structure emulates biological neural systems present in human and animal brains, whose synchronised use of multiple but simple processing units is capable of performing complex tasks such as learning behaviours, classifying data and be adaptive to environmental changes [109]. The processing units in these networks are *artificial neurons* while the connectors between units are named *synaptic weights*, similar elements to those present in the human brain (see Figure 2.7), discovered by Cajal at the beginning of the 20th century. [114].

While the technical jargon used in these networks differs from the one used by other systems identification methods, the components depicted in ANN are analogous to the latest in practice [88]. The first ANN proposed a binary threshold unit called the McCulloch-Pitts neuron so that the output y at each unit could be 1 or 0, depending both on the threshold and the value of the inputs x_i multiplied by the weights w_i associated to each (Equation 2.6).

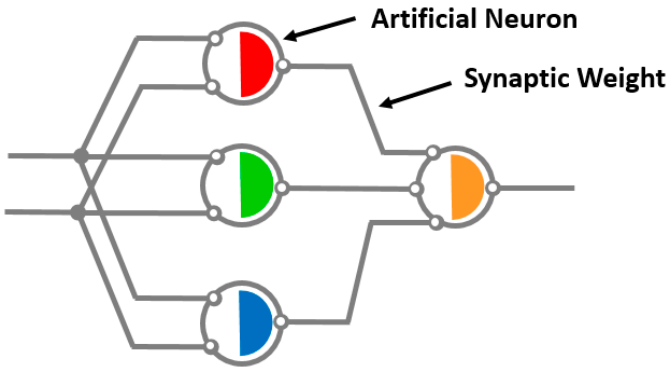


Figure 2.7: Basic components of ANN.

$$y = \varphi\left(\sum_{j=1}^{n} w_i x_i\right) \tag{2.6}$$

where $\varphi(\cdot)$ is a threshold function. Concerning the network structure, ANNs separate into parallel processing layers composed of many identical and interconnected neurons whose connection strength is defined by weights. Modern ANN learning algorithms iteratively modify such weights during the training process to minimise the difference of the model output concerning the desired output or pattern. This training enables a network to become a model made up by neurons getting involved in different degrees aimed at making the inputs to yield output values as closest to the actual values or expected answers (Figure 2.8) [88].

The most popular learning practice in ANN is the backpropagation method (BP), which has continuously been adapted and enriched since its first use in this context in 1982 [115]. When the ANN training adopts the BP algorithm, these are known as feedforward backpropagation neural networks (FFBPNN). The idea

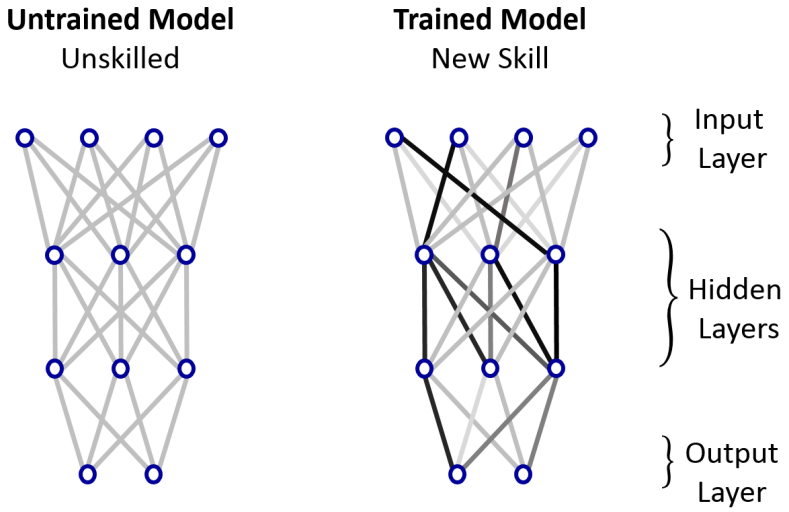


Figure 2.8: Learning process of ANN.

behind this process points toward the use of multiple *cycles* or *epochs* composed of:

(A) A forward propagation from input to output to obtain an approximation of the target value.

(B) Calculation of the approximation error concerning the target value.

(C) An error backward propagation from output to input to distribute the approximation deficiency throughout the weights connecting the neurons.

(D) Update of the network weights with an improved value to start the next epoch with an enhanced starting point.

The most used method to calculate the new weights minimising the error at step (B) is the gradient descent algorithm, which helps to indicate the direction (or plus-minus sign) in which the value of the weights are to be modified. Figure 2.9 shows the BP information flow.

Regarding the advantages, ANNs have proven to be effective at:

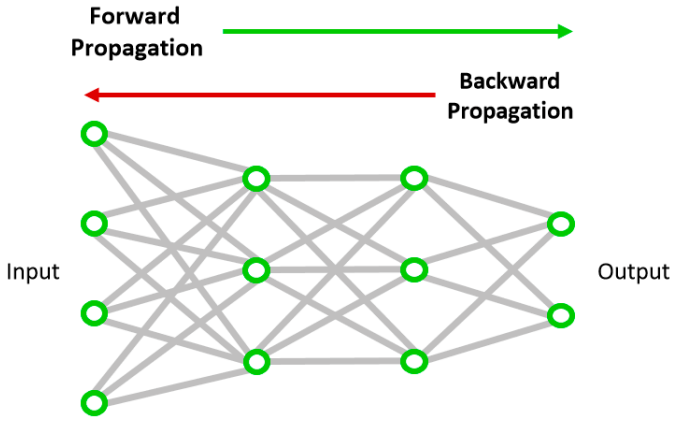


Figure 2.9: Types of propagation in ANN.

- Fast hardware implementation if single neurons can be put into effect efficiently.
- Improved tolerance to partial network faults thanks to parallel processing.
- Solving of nonlinear problems.
- Generalization capacity from known to unknown problems.
- Automated learning (no mandatory need of domain expert or task-specific programming).
- A fast operation after learning.

Although the mentioned points make ANN an appealing choice, these can present difficulties such as costly processing times. The last problem is because when the ANN has a large number of neurons or hidden layers, the model training task becomes more and more challenging. ANN have also a difficulty for modelling noise and they need for the availability of a large amount of training data, especially when the network is large [116],[88].

As regards applications, neural network architectures are recognised as functional modelling options pattern classification and clustering, given their competence to categorise and recognise different classes without previous labelling thanks their efficient learning algorithm, which is designed to work for a wide range of problems [117].

In object classification, ANN explores and assign, one by one, input patterns (in the form of feature vectors) to classes, according to their similarity. The general overview of this recognition process contains segmentation, mining of features, and contextual processing, which solution can require diverse ANN designs, including

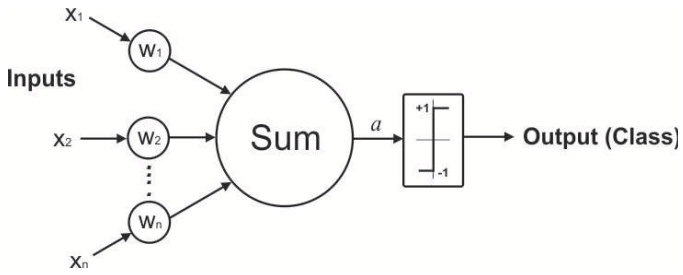


Figure 2.10: McCullough-Pitt artificial neuron model.

specific network structure and size, training sampling and training data, and neural activation function. While some authors affirm that the performance of ANN is very similar to the statistical approach, given that the first stems from the second, ANN allows combining different approaches and flexible techniques to reach good problem solutions [118]. Below are described significant variants of ANN.

- *Adaptive Neural Networks.* Adaptive neural networks are generally used as classifiers with complex decision boundaries, offering an improved generalisation and learning capacity. Adaptive networks offer higher parallelism, little energy expenditure and increased adaptiveness [119]. Adaptive NNs are multi-layer models which evolved from the first neuron prototype, which had a very similar structure to that of the linear discriminator classifier. However, in adaptive NN the output is fed to a nonlinear threshold function (Figure 2.10).

  Like the linear discriminator, this model can identify only two classes, unless many neurons are arranged simultaneously to obtain multiple patterns or classes. In [120], adaptive NNs are integrated into a CAD system to detect and classify lung nodules. Adaptive networks are first used to detect a series of suspicious zones in the image and then to classify each zone. The architecture of the adaptive NN based classifier includes multilayer perceptron archi-

tecture with two hidden layers, where the term *perceptron* alludes a higher ability to perceive patterns [32]. In parallel, the training of the adaptive NN classifier used the most typical patterns. Though the experimental tests showed a high detection effectivity, the method needs a further comparison to other approaches.

- *Deep Feedforward Networks.* Deep feedforward Networks (DFN) architectures represent the central archetype of deep learning models, which are part of machine learning and ANN methods [105]. Deep learning (DL), like ANN, use many simple units of mathematical processing inspired by biological processes with the ability to learn, as a whole, complex functions.

  DL is also a robust machine learning technique designed to extract features from input objects to automatically avoid human intervention during the learning process, adding in exchange more hidden layers for pre-training. This innovation reduced up to 20 times the previous training times for the speech recognition problem by 2009 [121]. Concerning the network architecture, the feedforward term of the DFN is due to the direction on which data runs through the network (Figure 2.11).

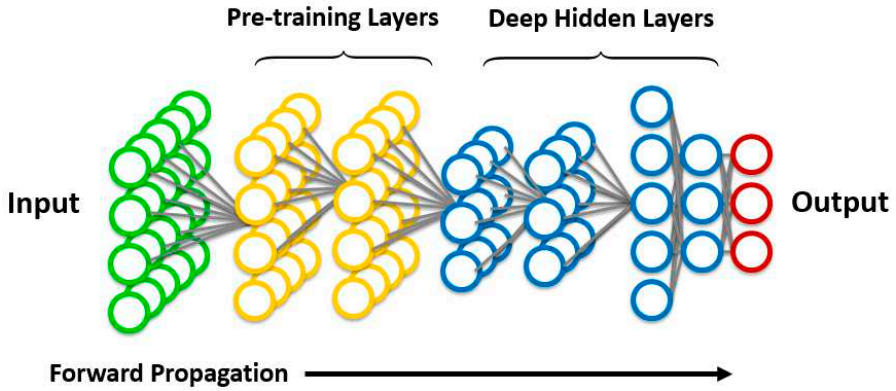


Figure 2.11: Structural example of DFN (adapted from [122]).

Although this network architecture is designed to progressively improve the quality of the input information with the aim of delivering sensitive feature values to the fully connected deeper layers of the network, the backpropagation training (detailed earlier in this section) is one of the most used learning strategies for DFN. This popularity is because BP contributes to adjust the deep network from top to bottom concerning the target value, principally in cases where the volume of training data is not massive [123].

In spite of the notable advantages, deep neural networks demand large data, high computational cost and a careful selection of their size given that they present loss of generalisation capacity when there are more layers than necessary [124],[125]. The main applications of the DFN architecture and deep learning models include computer vision in the form of visible sound waves recognition [126], denoising [127], contrast normalization [128] and dataset augmentation [129].

The network architecture most used in image processing problems takes the form of convolutional neural networks (see below). More applications include speech recognition, as in [130] where the recognition rate for the best-known topic-related database (TIMIT) improved by 6%, [131] where large vocabulary sentences were recognised and [132] where took place information alignment of acoustic and phonetic levels. With regard to natural language processing, there are examples in machine translation applications [133],[134], language parsing [135] and multitask learning architecture [136].

- *Convolutional Neural Networks.* Convolutional neural networks (CNN) [137] are conceived as ANN with the difference of containing at least one convolution layer within its architecture [105]. CNN are generally inspired by brain cells as well but particularly by the animal visual system [138]. For this

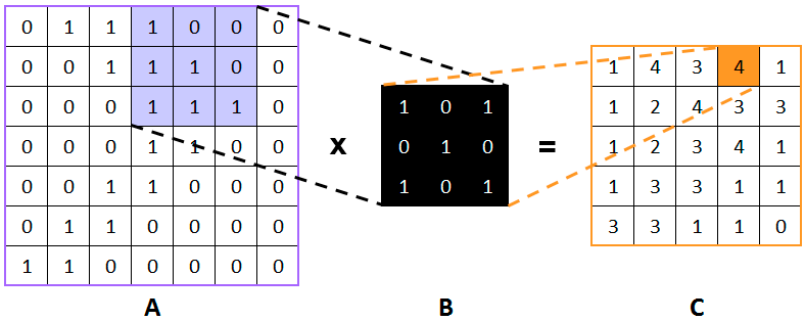


Figure 2.12: Convolution mask processing over an image.

reason, the main application of CNN is image understanding and classification, although there are additional applications to be seen later on in this section.

The initial stage of this networks builds upon convolutional and pooling layers working in tandem, where the first ones aim to filter the input through a linear mathematical process termed *convolution* while the former ones seek to merge similar features into single representative values [123]. That filter arises from the *convolution kernel* definition, that estimates several parallel convolutions to interpret and decode the input grid into space-referenced feature maps.

Another way to understand the logic of the convolution function is that it provides a polished estimate of the multidimensional input arrangement to acquire the most relevant values to be taken into account subsequently by the fully connected hidden layers. Figure 2.12 exemplifies the use of a function of this kind, where image A is convoluted by kernel B with a resulting compacted grid C.

After using the convolution layers, the feature maps travel to a subsequent intermediate layer known as *pooling layer* that helps to downsample or reduce

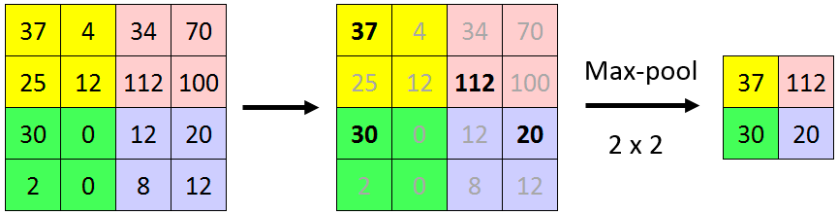


Figure 2.13: Max pooling for a 2×2 neighbourhood.

the data dimensionality before being processed by deeper network layers [105]. The most common pooling function is the max-pooling [139] that operates by selecting the maximum value of a neighbourhood of $n \times n$ size (Figure 2.13).

CNN represent a significant and growing field in computer vision. Remarkable applications encompass image classification [129] where a deep CNN was used to reduce the recognition error of a big database by nearly half, human pose estimation [140] where image recognition processed RGB images, handwritten recognition [141] where a backpropagation training was carried out for the first time in a convolutional network and speech recognition [142] where a time-delay neural network and error backpropagation were combined.

More applications include optical character recognition [143]] were the method was bought by Microsoft time after, text modelling [144] where CNN and recurrent NNs processed sentence modelling and classification, face verification [145] in which several interleaved convolutional and pooling layers manage to extract high-level biometric features, and action recognition [146] in which a deep convolutional mapping characterized video representation focused on trajectory extraction.

- *Extreme learning machine and autoencoders.* Neural networks are regarded as universal approximators of the unknown function relating the inputs and the outputs of a nonlinear system [88]. As seen earlier, it is common that the learning in neural networks takes place via backpropagation and the gradient-descent algorithm, the latter being by far one of the most common nonlinear optimisation methods and the most popular machine learning algorithm used in neural networks [147].

  However, the mapping of a model with this method has a generally slow convergence rate which may last several hours or even days, which is a problem in practical cases such as those in the industry [148]. Also, solution algorithms based on gradient-descent run the risk of producing overfitting and falling into local minimum [149].

  The extreme learning machine (ELM) [150] is an unsupervised and much faster feedforward NN since on the one hand it is not iterative and propagates in a feedforward direction and on the other, it leaves out the gradient-descent approach. The ELM bases its speed in a random process to determine the weights connecting the inputs with the hidden layer since test results show that the network training can do well without the standard weight estimation [151]. Nonetheless, to estimate the weights connecting the hidden layer to the output, the ELM uses a pseudoinverse learning process.

  An essential capability of the ELM is the functionality for feature extraction and dimensionality reduction or compression, especially when coupled to autoencoders (AE) [152]. AE have a unique network architecture with a compacted or bottleneck-type hidden layer aimed at duplicating the system input and place it in the output as a pattern. In that architecture, the weights connecting the intermediate layer to the last one are forced to represent

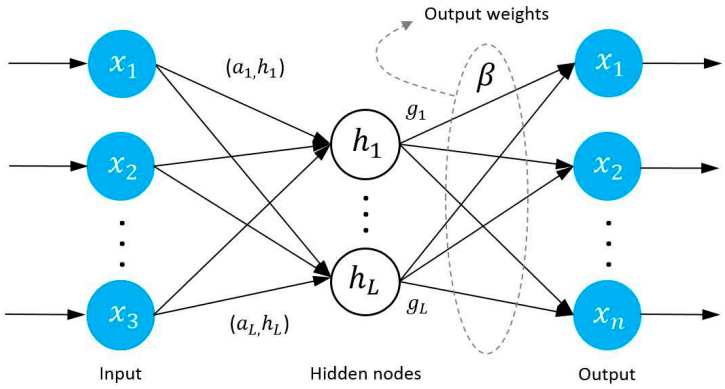


Figure 2.14: ELM-AE architecture (taken from [153]).

the dimensionality change, mapping at the same time the extracted features according to the network architecture [153].

Another practical idea behind the ELM-AE architecture is to extract a set of high-quality features to be understood and processed by a final classification layer. The basic ELM- AE structure for feature extraction is highly efficient and has proven to be suitable both for kernels and deep neural networks structures. Figure 2.14 illustrates the scheme.

The definition of the ELM-AE unsupervised learning algorithm, as defined in [154] is as follows:

$$f_{ELM}(x) = \sum_{i=1}^{L} \beta_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} \tag{2.7}$$

where $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_L]^T$ are the weights connecting the hidden layer and the output, $\mathbf{h}(\mathbf{x}) = [g_1(x), \cdots, g_L(x)]$ are the outputs values of the hidden nodes after processing the input $\mathbf{x}$. With the previous definition, the ELM-AE

learning problem remains in the computation of the output weights $\boldsymbol{\beta}$, which can be written in the matrix form as:

$$\boldsymbol{\beta} = \mathbf{H}^{\dagger}\mathbf{T} \tag{2.8}$$

in which $\mathbf{T} = [t_1, \cdot, t_n]$ is the vector of target values, also known as patterns or labels, $\mathbf{H} = [\mathbf{h}^T(\mathbf{x}_1), \cdot, \mathbf{h}^T(\mathbf{x}_N)]^T$ and $H^{\dagger}$ is the pseudoinverse of $\mathbf{H}$. Then, the ELM pseudoinverse learning problem is defined as:

$$\boldsymbol{\beta} = \mathbf{H}^T(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T)^{-1}\mathbf{T} \tag{2.9}$$

where C is a regularisation parameter to avoid overfitting. Finally, the solution of $f_{ELM}(x)$ comes about by replacing Equation (2.9) in Equation (2.7).

The ELM and AE networks for feature extraction have been reported in different formats and tested in many classification problems. Among the most relevant, in [154] the ELM and AE were put together for the first time into a deep neural network termed ML-ELM. The method was tested in the public MINST dataset and compared to prevailing deep network structures. The ML-ELM ran significantly faster than common AE and had superior or similar classification performance than the selected benchmarks.

In [20], a kernel version of the multilayer ELM was proposed along the adoption of AE networks to make the information to advance through the network from one stacked layer to another without the need for human supervision. In spite of being more straightforward, the ML-KLEM network was proven to be much faster than two previous multilayer networks. The classification accuracy for 20 public datasets for different problems outreached the other methods in all cases, especially in large datasets.

In [153] a generalised version of the ELM-AE was reported by adding an improved regularisation process. Besides, the authors used a stacked AE deep NN architecture, which is capable of repeating and improving the feature extraction process several times through the network structure. Tests with 13 public image sets showed that the GELM-AE improved the classification accuracy in almost all cases versus classic and deep learning classification benchmark methods. As for the breast cancer classification problem, Section 2.4 reports related methods.

Radial basis function networks

Radial basis functions (RBF) are a special kind of artificial neural networks in which there is only a single hidden layer, unlike deeper configurations known as deep forward networks, reviewed before in this section.

Within the context of machine learning and system identification, this type of networks were proposed in [104] with the aim of approximating an unknown function from known data by summing several identical basis functions aimed at making up an intermediate layer of the neural network known as *hidden layer*. Figure 2.15 shows the typical architecture of an RBF neural network.

This network establishes a relationship of the system output y with the input vector $\mathbf{x}$ through an unknown nonlinear function:

$$y = f(x_1, x_2, ..., x_n) = f(\mathbf{x}) \tag{2.10}$$

To develop the function each neuron of the intermediate layer is regarded as a basis function φ known as *kernel*, which intends to provide a measure of similarity (typically the distance) within a multidimensional space between a sample and a centre through their internal product. The single-hidden-layer of an RBF network stands as the sum of kernels and aims at approximate the unknown function $f(\mathbf{x})$. The following formula expresses the above:

$$\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{M} w_j \varphi_j(\boldsymbol{x}) \tag{2.11}$$

where φ_j represent the radial basis function for the jth neuron and w_j the corresponding weight. The most used kernel in RBF is the Gaussian one, with the Euclidean distance as the norm:

$$\varphi_j(\boldsymbol{x}(t); \boldsymbol{\sigma_j}, \boldsymbol{c_j}) = \exp\left[-\frac{1}{2}\left(\frac{x(t) - c_j}{\sigma_j}\right)^2 \right] \tag{2.12}$$

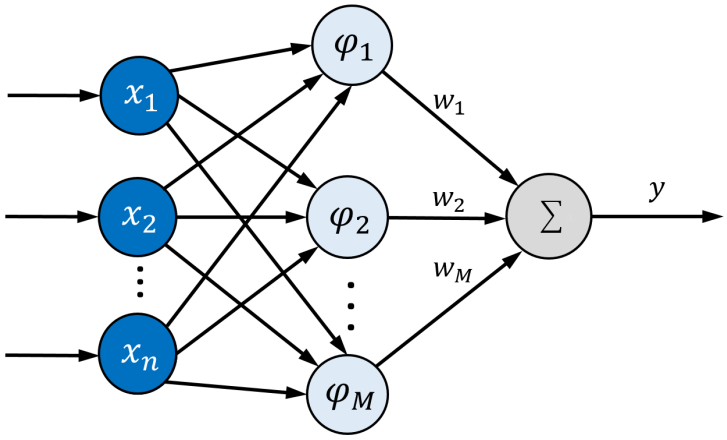


Figure 2.15: Basic scheme of an RBF neural network.

where σ_j and c_j are vectors containing the scales and the kernel centres for the jth neuron. Although this category of neural networks is the simplest one concerning hidden layers number, in [155] and [156] and later in [157] was proven that the utilisation of this layout setup is sufficient to model any nonlinear function. Also, RBF networks simplify the nonlinear functions approximation by reducing the determination of the weights (θ's in Equation 2.3) to linear expressions [104].

For this reasons RBF networks converge faster and adapt more favourably during the learning process, especially in problems of classification where the training set is sufficiently small [116]. Although the mentioned points are advantageous and produce computationally easier training, special attention must be placed in the determination of the quality and quantity of the kernel centres $\boldsymbol{c_j}(x)$, as these represent a starting point to the training or approximation algorithm.

RBF networks have been used successfully in several contexts. In [15] a study for modelling and predict the reactivity of near-earth geomagnetic field to magnetic storms is reported. The authors took two solar wind-related variables as inputs and the resulting disturbance in the magnetosphere as output. The modelling introduced multiscales to give the RBF network greater description flexibility for non-linear systems. The forward orthogonal regression (FOR) algorithm (close related to FROLS, detailed later in Section 2.2.3) was adopted in the identification structure to simplify the problem into a linear-in-the-parameters form.

Other studies include the modelling and identification of dynamical systems, as in [158] where an RBF network competed against multi-layered networks for solving different problems. In general, the study confirmed the excellent capacity of neural networks to represent nonlinear systems. In particular, the capability of RBF networks to derive linear learning laws show them to be faster of converging and more accessible to train in identification problems.

Below are more examples of applications of this networks in image classification. In face recognition, authors of [159] found that a small sample set is enough to train the network for a classification process successfully. The work in [160] recommends three-dimensional object recognition where the learning process proceeds from a small image set composed of projected views of the object of interest. Authors of [161], propose a motor systems control where the parameters of the RBF work as optimal values of a velocity sensor. In medical image analysis, the works [162] and [163] use the RBF network in each case as a classification tool, after making a process of image decomposition into feature vectors.

In this way, it was possible to identify the pathological samples from the healthy ones with competitive results in both mammography [162] and brain images [163]. Note that the term *pathological* refers to the state of a person suffering or being affected by a disease [164], [163]. Section 2.3 reviews additional system identification approaches for breast cancer detection different than RBF.

Fuzzy logic models

Fuzzy sets emerge in 1965 as an effort to extend the precise (crisp) quantitative analysis into qualitative and uncertain problems which are generally faced by most real-life human-related disciplines such as social, medical, political and economic sciences. Under this logic, the notion of fuzzy set refers to an object that contains elements with partial degrees of membership about certain concepts. Fuzzy logic models are also mappings with established rules for input-output systems of the general form $y = f(x)$. This kind of modelling grasps four processing modules: establishment of rules, fuzzification, inference, and defuzzification (output processor) [165].

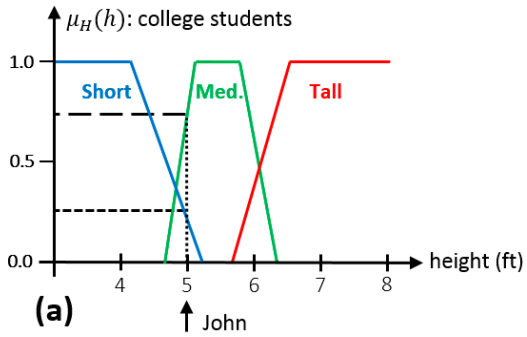


Figure 2.16: MF of average students.

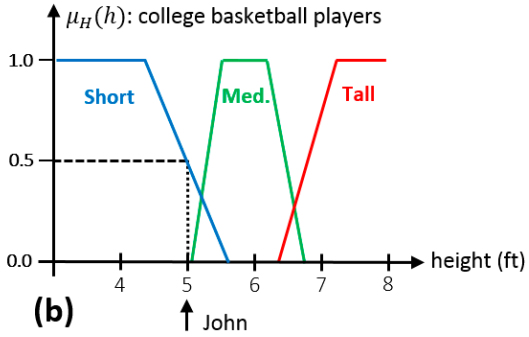


Figure 2.17: MF of taller students.

A key concept to take the fuzzy logic into practice is the *membership function* (MF), which is a suitable representation mechanism capable of mapping and quantifying the conceptual complexity occurring in problems with uncertainty. The MF, as a distinctive feature of fuzzy logic models, associates *shades* of different classes to objects.

To illustrate the concept of MF, below are shown two instances in which the degree of membership μ_H for the measured variable (height) changes for two different student groups; Figure 2.16 for male college students and Figure 2.17 for male college students who belong to the basketball team (example adapted from [165]).

The graph on the left indicates that John, a 5 ft. tall student, belongs to the *short* set in a 0.25 degree, 0.75 to the *medium* set and 0 to the *tall* set. In the case on the right, the same student belongs to the *short* set in 1.0 degree and 0 to the other sets. This example makes simpler to see that the height value (short, medium, tall) has a different meaning in different contexts.

Fuzzy sets can be either type-1 or type-2 depending on the uncertainty partition order the set is described [166]. To explain such a difference, we should consider that the observed variable (horizontal axis in Figures 2.16,2.17) splits up

into intervals where the existing classes overlap. The first-order uncertainty allows knowing the precise points where the overlapping begins and ends. In second-order uncertainty, however, this one-dimensional precision is removed in the service of improving the mapping of uncertainty in the problem.

Another difference is that in type-1, the membership degree for a value of the measured variable within any class overlap is a real number that exists in $[0, 1]$. On the other hand, the type-2 membership degree for the same variable value represents an interval of real numbers which is a subset of $[0, 1]$ [165]. More formally, type-1 fuzzy sets are:

$$A = \{(x, \mu_A(x))|x \in X\} \tag{2.13}$$

where A is a set function contained in the universe X and $\mu_A(x)$ is the MF of A, where $0 \leq \mu_A \leq 1$.

Conversely, type-2 MFs can be seen as the blurred version of the type-1 MFs. The type-2 MF depends on two variables (x and u) and is represented as $\mu_{\tilde{A}}$, where $\tilde{A}$ is a type-2 fuzzy set, and $0 \leq \mu_{\tilde{A}} \leq 1$. Then, the type-2 fuzzy set is defined as [21]:

$$\tilde{A} = \{((x, u), \mu_{\tilde{A}}(x, u))|x \in X, u \in U \equiv [0, 1]\} \tag{2.14}$$

where $\tilde{A}$ is a set function contained in X, the universe for the variable x, and U, the universe for the variable u. The *footprint of uncertainty* (FOU) is the type-2 version of the type-1 membership degree. The FOU represents a region of the Cartesian product $X \times \{\mu_{\tilde{A}}(x)\}$ into [0, 1], where the membership degree $\mu_{\tilde{A}}(x)$

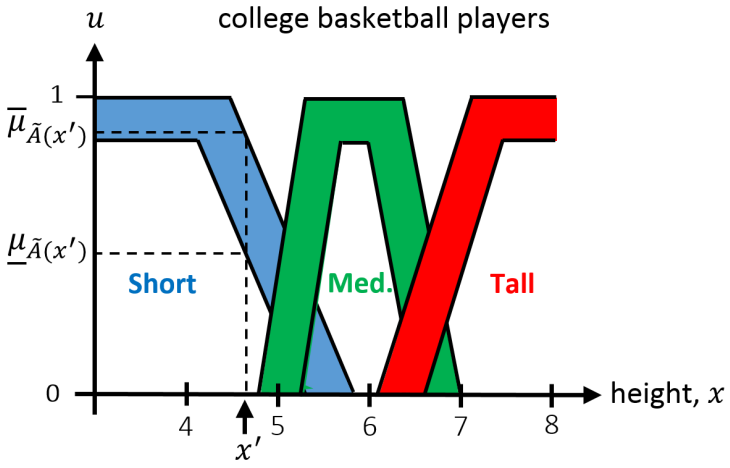


Figure 2.18: Three FOUs for the variable *height* (adapted from [165]).

is an interval of real numbers instead of a single value. Bearing these concepts in mind, the MFs of Figure 2.17 can be reinterpreted as FOUs, as Figure 2.18 shows.

In the example, $\underline{\mu}_{\tilde{A}}(x')$ and $\overline{\mu}_{\tilde{A}}(x')$ are the upper and lower membership functions which form an interval of values (subset of [0, 1]) brought by the FOU for the type-2 fuzzy set $\tilde{A}$.

Applications of fuzzy logic models often include hybrid models for diverse purposes as classification, diagnosis, searching, evaluation, decision making, control and planning. The work in [167] proposed a classification tool for breast cancer and heart disease through a hybrid model of fuzzy logic and a genetic algorithm for high dimensional data. Results showed that the fuzzy component of the model helped the method to process uncertainties positively. In [168] fuzzy search was implemented with a conditional random field in a text-mining technique for disease name recognition for biomedical literature. While the conditional random field served as a base classifier, a fuzzy search was employed to label unusual disease names.

In [169] a fuzzy-based decision evaluation method in curtain grouting was proposed. Curtain grouting is the hydraulic barrier under a dam to reduce water leakage. The fuzzy assessment took into account permeability, rock quality and tightness of the rock mass to determine the best execution plan. The method recommended an efficient strategy and led to a better understanding of the problem. In [170] a fuzzy system was presented to tack the trajectory of marine vehicles, a problem with uncertainties as currents and waves. The adoption of a structure learning mechanism helped to create automatically easy to interpret fuzzy rules and fuzzy sets capable of identifying uncertainties. Results showed that tracking performance improved previous methods for a similar instance.

In [171] an RBF-fuzzy granular approach for modelling problems with uncertainties was put forward. In the preprocessing, the method applies the concept of granular compression to compact the system inputs into a finite number of granules which group similar data, leaving out most uncertainties. Next, the RBF network hyperparameters and membership functions are calculated from the granules to build up the RBF-fuzzy model.

The proposed modelling framework addressed a case of study with uncertainties known as the Charpy impact test, which measures the strength of materials by applying external stress. Results showed that in spite of the high uncertainty of the problem, the method improved the testing performance of previous models as regards generalisation capacity and global accuracy for predicting the strength of materials.

2.3 Detection of the model structure

Once the unknown function $F[\cdot]$ has been defined in the form of a nonlinear structure according to the model of choice, a crucial role of the identification process comes into play: the selection of the correct terms or regressors to include in the final model. More than often circumstances do not make easy for operators to have information revealing the structure of a model intended to describe the phenomena, so it becomes necessary to adopt suitable algorithms to detect it. This necessity includes the ability to identify as simple as possible models without sacrificing representability until finding, if any, the most basic rule connecting input and output values.

For instance, this advantage would prevent getting a nonlinear model for a simple linear problem. Structure detection faces other challenges as the restricted availability of user-friendly toolboxes, requirement of preventing long training times and a limited amount of algorithms for solving dynamic models [109]. Figure 2.19 shows the general flow of the structure detection process.

The most important model structure detection algorithms for nonlinear system identification include term clustering [172], multi-objective error reduction [173],[174], forward orthogonal least squares [175],[176], evolutionary algorithms [177],[178], local linear model trees (LOLIMOT) for fuzzy models [179], least absolute shrinkage and selection operator (LASSO) [180] and heuristic optimization [181],[182].

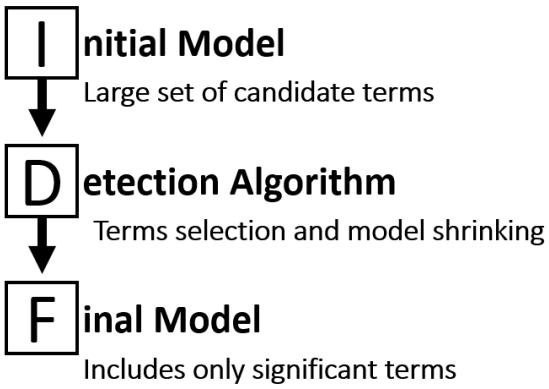


Figure 2.19: General strategy for detecting the model structure.

2.3.1 The FROLS algorithm

The Forward Regression Orthogonal Least Squares (FROLS) algorithm [183],[184], known as well by the name of orthogonal forward regression (OFR), is a widely used algorithm for structure detection of nonlinear systems. It adopts the advantages of the Orthogonal Least Squares (OLS) algorithm [183] working together the Error Reduction Ratio (ERR) estimator [185] and adds a reordering process which gives the joint algorithm more efficiency [183].

In the basic OLS, a first objective is to transform the original linear-in-the-parameters representation of the regression model into another one with mutually orthogonal regressors. Then, the new representation is used by the ERR algorithm to produce an iterative term selection. The initial linear in the linear-in-the-parameters representation is:

$$y(k) = \sum_{i=1}^{M} \boldsymbol{\theta}_i p_i(k) + e(k) \tag{2.15}$$

where $y(k)$ is the output sequence, M is the total number of candidate terms, $p_i(k)$ is the sequence of model regressors (namely candidate terms) made up by combinations of input and output variables contained in $\boldsymbol{x}(k)$ (see Equation 2.4), $\boldsymbol{\theta}_i$ are the model parameters and $e(k)$ the error sequence. From the matrix perspective, the Equation (2.11) can be represented as:

$$Y = \boldsymbol{\theta} P + e \tag{2.16}$$

whith P standing as the matrix of model regressors. Thus, an operation known as the QR decomposition [186] produces an orthogonal break down of matrix P into W and A:

$$P = WA \tag{2.17}$$

in this way, P can span into an M-dimensional vector subspace, A is an upper triangular matrix and W is a matrix with orthogonal columns $\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_M$, where each column is an orthogonal basis, froming a basis set of the same size as the subspace spaned by P. This process, based on linear algebra, is the Gram-Schmidt orthogonalisation algorithm [187].

The OLS algorithm takes advantage of the previous procedure and uses the ERR to estimate the contribution of each candidate term i (with $i = 1, 2, \cdots, M$) to the variance reduction with respect to the desired output $\boldsymbol{y}$ by using the mutually orthogonal basis contained in W. More formally, the ERR is stated as follows:

$$ERR_i = \frac{\langle \boldsymbol{y}, \boldsymbol{w}_i \rangle^2}{\langle \boldsymbol{y}, \boldsymbol{y} \rangle \langle \boldsymbol{w}_i, \boldsymbol{w}_i \rangle} \times 100 \tag{2.18}$$

where $\boldsymbol{w}_i$ is an orthogonal vector corresponding to the ith candidate term, $\boldsymbol{y}$ is the vector of the desired output, from which deviations are measured, and the notation $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors.

As part of a joint algorithm, the ERR helps to list the candidate terms in order of significance to the deviation reduction concerning the output, thanks to the ERR values linked to each possible choice. Thus, the structure detection algorithm can select and include term by term the most significant elements in the solution model. The process goes on until the model's deviation concerning the output reaches a threshold, pre-established with respect a minimum desired accuracy. The latter can, therefore, be seen as stop criterion which is verified each time a new term is selected and included in the model throughout the error-to-signal ratio (ESR), which computes the model accuracy by summing the ERR values linked to the selected terms. The detection algorithm calculates the difference between such summed value (ESR) and the desired accuracy threshold and stops the iterations when the former reaches the latter. The formula of the ESR is:

$$ESR = 1 - \sum_{i=1}^{M_s} ERR_i \tag{2.19}$$

where M_s represents the number of candidate terms selected in the final model so that $M_s < M$. Please note that for each iteration, the Gram-Schmidt orthogonalisation algorithm [187] guides the detection algorithm to exclude from the final model the candidate terms providing redundant information to that given by the terms already included in the model. The OLS combined with the ERR has proven to be superior to the least squares algorithm thanks to the orthogonalisation process that yields an improved exclusion of the redundant terms [185].

Please also note that the ERR and the OLS are complementary in this approach since in each iteration the former assigns the contribution value of each candidate to the error reduction concerning the output $\boldsymbol{y}$. Meanwhile, the OLS helps to relegate from the final model the candidates with repeated information to that of the candidates selected in previous iterations. However, this combined approach can mistakenly confer higher ERR values to the regressors of $p_i(k)$ appearing first in the Equation (2.11), producing a partially influenced term selection process [188].

The FROLS algorithm uses the OLS and the ERR algorithms and solves the ordering problem. The solution takes place by introducing a simple but efficient reordering technique in the terms comprised within $p_i(k)$, which is the pool of regressors to be orthogonalised by the Gram-Schmidt algorithm [187] during the OLS. Such reordering, described in [88], involves the following steps:

1. The reordering takes into account the lower order terms firstly, which are usually linear.

2. For each order, the regressors containing more y and less u variables are placed first in the equation. For instance, in the second order iteration, the term $y^2(k-1)$ would be placed first because it does not contain any u variable.

3. Following this, in the middle are placed the regressors with a balanced number of y and u variables. For example, $y(k-1)u(k-1)$.

4. The regressors that contain more u variables and less y variables, e.g. the term $u^2(k-1)$ would be placed by the end of the formulation.

5. Afterwards, the inclusion of the higher order terms takes action through steps 2 to 4.

6. The process continues until the higher order terms are rearranged. In this way, the global structure detection algorithm can perform the candidate selection without incorrect biases.

The FROLS algorithm has been successfully used as structure detection algorithm in numerous nonlinear approximations. Notable examples include radial basis functions [189],[190], fuzzy systems [191], neural networks [158],[114] and sparse models [192].

2.4 Computer aided diagnosis

Computer-aided diagnosis (CAD) systems represent a central research branch of medical image processing. The accumulative increase of hardware and software plus the access to advanced and new image processing techniques put CAD technologies within an unmatched perspective in its history. Initially, CAD arose as an attempt to replace radiologists by computers [193].

However, the new CAD philosophy does not search to replace physicians, but to offer instead a fast, objective and accurate *second opinion* to detect a variety of anomalies in early stages ranging from lung nodules, vertebral fissures, size of hearts and malignant nodules and prevent unnecessary biopsies, the proliferation of cancer or its development to more advanced stages [194],[34].

CAD procedures emerge from multiple knowledge areas including biosignal processing, digital image analysis and statistics. The ultimate purpose behind this mixture is to provide a reliable inference about a potential health disorder, based on a sophisticated extraction of information from measured biosignals, including medical images [194]. Figure 2.20 shows a general flowchart of this process.

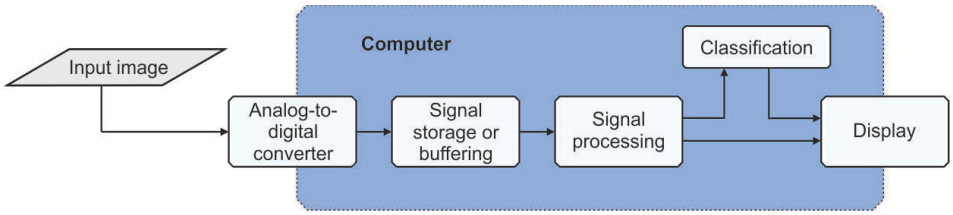


Figure 2.20: Flowchart of medical image processing and CAD systems. Taken from [32].

2.4.1 CAD for breast cancer

Breast cancer is a severe public health problem in many countries, regardless of their development level. Only in the United States during 2017, an estimated of 40,610 women died of this disease, while another 63,410 new cases of breast carcinoma in situ (a disease in early stage) were diagnosed, representing 30% of cases of cancer of any kind in women [23]. In developing countries, breast cancer has become the leading cause of cancer death among females, contrary to the past decades [195]. While medical science has found relevant findings regarding the causes and treatment of the disease, early diagnosis remains as the best current strategy to improve the prognosis and treatment for affected women. CAD systems have received attention because of its increasing power to assist humans at finding abnormalities and discriminate malignant from healthy tissue regions within medical images, either with assisted or unassisted learning processes. CAD systems are meant to extract specks, blobs or distinguish suspicious from healthy regions from digital mammograms in breast cancer diagnosis [196].

2.4.2 Diagnosis performance metrics

The basic and most common performance metrics found in the literature on CAD systems for digital mammography are accuracy, sensitivity and specificity, while positive predictive value (PPV) and negative predictive value (PNV) are found with less frequency [197]. More specifically, accuracy is a straightforward metric but it ignores the disease prevalence. Sensitivity and specificity do not make this omission and quantify how consistent a classifier is to not overlook (fail to notice) positive and negative cases, respectively. PPV explains the chance that a case defined as positive is unhealthy. NPV calculates the chance that a case classified as negative is healthy. The corresponding formulas are below ([198]):

$$Accuracy = \frac{Correct\ decisions}{No.\ cases} \times 100 \tag{2.20}$$

$$Sensitivity = \frac{True\ positive\ decisions}{Actually\ positive\ cases} \times 100 \tag{2.21}$$

$$Specificity = \frac{True\ negative\ decisions}{Actually\ negative\ cases} \times 100 \tag{2.22}$$

$$PPV = \frac{True\ positives\ decisions}{True\ positive + False\ positives} \times 100 \tag{2.23}$$

$$NPV = \frac{True\ negative\ decisions}{True\ negatives + False\ negatives} \times 100 \tag{2.24}$$

2.5 System identification models and ANN into CAD for breast cancer

In recent years an abundant collection of system identification parametric models and artificial neural networks have been presented as a feasible alternative for medical image analysis in the detection of breast cancer. These efforts, all within the CAD paradigm, have been directed to process digitalised breast imagery coming from X-rays, ultrasound, magnetic resonance, and electrical impedance tomography. The work considered as the most relevant and representative in this subject are listed below.

2.5.1 Parametric model-based CAD systems

In [199] a one-dimensional autoregressive (AR) filter was proposed to improve the contrast of ultrasound breast cancer images automatically and to enhance the visualisation of suspicious masses. The approach used the concepts of higher harmonics of the frequency band and characteristics of the sound wave propagation in compressed versus relaxed tissues. The work inferred the presence of potential malignant tissues through 1D signal and image contrast enhancement. The model, though one-dimensional, was able to represent accurately real-life data.

In [10] an ARMA model to detect and classify breast cancer was proposed as an innovative CAD scheme for ultrasound images. This approach, unlike previous 1-D methods [200],[201] and [199], presented a 2-D image analysis for classification by using a moving window scan to take into account the spatial correlation. The Yule-Walker least-squares algorithm computed the model parameters, while the K-means classifier used the ARMA parameters as feature vectors to diagnose

the breast image as healthy, benign or malignant. Simulations with real medical databases proved a general accuracy above 90%. However, the lack of 2-D models available for comparison motivated the authors to compare versus the 1-D ARMA model only.

In [200] and [202] two FARMA CAD modelling for one-dimensional tissue characterization from ultrasound radio-frequency echo schemes were presented. The main modelling assumption stemmed from the observed similarity between ultrasound echo frequencies reflected by tissue and fractal processes. In consequence, the incorporation of fractional parameters F in the ARMA models allowed to capture the fractal parameters. Experimental results confirmed that the best FARMA configuration attained an accuracy of 87%, overtaking by 6% the radiologist's prebiopsy criteria. Although these methods were limited to ultrasound images, the encouraging results suggest their use in other image analysis problems.

In [11] the same authors introduced the ARMA model for image feature extraction and the change detection algorithm based on sequential statistical analysis for microcalcification detection. In this work, the ease for solving the fixed-structure ARMA linear model was exploited to extract parameters from the model and statistically analyse abrupt changes in the parameters sequence, changes linked to the appearance of microcalcifications in the image. The tests showed that the method achieved sensitivity and specificity values above 94%. Despite good results, this approach works only with microcalcifications within the broad spectrum of tumour types. The work in [79] presented a similar CAD parametric approach for breast cancer for electrical impedance tomography.

In this case, an ARIMA model aimed to capture the image stationarity, understood as the addition of continuous regular processes, for instance, a repetitive pattern or a continuous arrangement along the image. Additionally, the 3-D multi-

frequency electrical impedance mammography (MEM) machine emerged as an affordable and non-invasive breast scanning option especially useful in developing countries. The method's main steps included image enhancement, a 3-D to 2-D image conversion, the Yule-Walker 2-D parameter estimation, and the K-means classifier. Although the accuracy of simulations was not specified, the authors claimed that the method is more effective than its 1-D counterpart, a more in-depth comparison versus other 2-D models could have taken place.

In [12] an autoregressive quantitative ultrasound characterisation (AR-QUS) model for breast cancer was presented. The study analysed images at a cellular level and compared directly the differences between tissue types, where an autoregressive model estimated the power spectrum of tumour data. The algorithm was capable of discerning healthy versus cancerous tissue but failed to distinguish the tumour type.

More recently, authors of [10] presented in [203] a 2D-ARMA model and a 1-D change detection algorithm as modelling strategies to detect small calcifications in mammograms. The addition of the change detection algorithm obeyed an observed link between calcification presence and statistical-additive changes in image's local properties. To model the tumour detection, the one-dimensional change detection algorithm took the probability density functions to characterise the image features obtained by the ARMA model. In this way, a series of additional parameters θ resulted from averaging the PDFs via the generalised likelihood ratio (GLR). After the parameter determination, a threshold value t_a was experimentally chosen and fixed to compare all parameter values to infer if the whole image resulted as abnormal. Simulations with 524 normal and cancerous cases showed that the accuracy surpassed 92%.

2.5.2 ANN-based CAD systems

In [162], a CAD system for breast cancer detection using the grey-level co-occurrence matrix for feature extraction and an RBF neural network as a classifier. Test results of the RBFNN compared to results of a back propagation ANN showed that the reported method came to be better in accuracy (93.9% vs.79.5%) and tumour class distinction (100% vs 89.5%). In [8] an easy-to-implement CAD approach used independent component analysis for feature extraction and RBFNN for classification to attain an accuracy of 88.2% and abnormality distinction rate of 79.3%.

In [9] an ANN technique for breast cancer detection was introduced by using a grey level co-occurrence matrix for feature extraction and the scaled conjugate gradient backpropagation to train the network. Classification results were positive for accuracy and sensitivity (93.1% of, 99%) but only moderately good for specificity (83%). In [204] an integrated CAD system for breast cancer detection using a particular network architecture was presented. The authors proposed the generalised pseudo-Zernike moment for feature extraction which is claimed to be robust to noise, and a new adaptive differential evolution wavelet neural network was recommended as a classifier. Two mammogram databases were used during testing (MIAS and DDSM) attaining accuracy rates of 89% and 87% respectively.

As for extreme learning machine and autoencoder networks in CAD, in [205], an ELM network-based CAD system for breast mass classification was reported. The framework included image segmentation as a preprocessing step to remove background artefacts in the first place. Then, the isolation-extraction of the regions of interest (a breast tumour and surrounding area) took place via the Hough transform. To compare the efficiency, ELM, support vector machine (SVM) and particle swarm optimisation plus SVM (PSO-SVM) where compared. Classifica-

tion performance results with a series of ROIs extracted from the MIAS database showed that the ELM-based method averaged accuracy of 95.73% from 5 different training- testing folds, which was notably superior to PSO-SVM (90.5%) and SVM (89.5%). However, there was no comparison to previous work results.

In [7] a CAD system using ELM for breast cancer detection was introduced for the classification of benign and malign tumours enclosed in regions of interest from the MIAS database. This unsupervised method aimed at speeding up the training and achieving better generalisation properties, while a subset of 9 out of 15 image features resulted from by using a heuristic search for better image representation. An accuracy of 91% in testing was attained by ELM, beating the metrics of previous approaches including SVM (82%), random forest decision classifier (90%) and K-nearest neighbours (84%). In spite of the advantages, the algorithm accuracy was proven to be dependent on the fair selection of the image features.

Authors in [18] reported an orthogonal incremental ELM autoencoder for image classification. It introduced the Gram-Schmidt orthogonalisation method coupled with Barron's convex optimisation learning algorithm to estimate the optimal weights connecting the hidden and the external layer of the network and simultaneously reduce the convergence rates, unlike previous incremental ELM. Testing on 5 classification problems for unidimensional data, including a breast cancer database of 1029 clinical cases, and 6 benchmark ELM-based methods revealed that the OCI-ELM testing accuracy outperformed 5 out of 6 methods in the breast cancer problem (94.73%). Additionally, tests with the MNIST image database of handwritten digits shown that OCI-ELM outperformed all benchmark methods with an accuracy of 97.89%.

In [206], the multiobjective optimisation of deep AE for feature extraction and dimensionality reduction was described and tested in mammogram classification. Multiobjective optimisation aimed at reducing both the reconstruction error between the AE input and output and the classification error to improve the whole feature extraction process. Tests for 949 mammograms using 8 different classifiers (1 at a time) in the network's classification layer showed that the accuracy of the optimised AE network reached classification accuracies from 80% up to 98.45%.

2.6 Chapter remarks

- System identification brings together techniques capable of describing and reproducing simple and complex, non-linear systems with flexibility and transparency.
- System identification is commonly applied in several real instances, especially in systems with inputs and outputs that are interrelated by an unknown process.
- System identification can be used along with other algorithms to solve different tasks such as analysis, modelling, replication of results, forecasting, risk prevention and classification,
- Image processing and computer vision are interdisciplinary fields that have reduced their cost in line with the advance of technologies to capture, broadcast and store digital information.
- Image processing takes elements from computer science and signal processing. Image classification uses statistical learning techniques.

- Popular image processing techniques include image segmentation, edge detection and object recognition, which gives this field a broad application landscape.

- Image processing can assists humans in several fields of application such as remote sensing, surveillance, manufacturing, robotics and autonomous vehicles, and sciences, like agriculture, meteorology and medicine.

- Image processing has produced relevant results in medicine as a tool to support the medical decision making in disease analysis, detection, monitoring, classification, diagnosis and prognosis.

- Medical image processing has a particular interest in breast cancer detection due to its worldwide incidence and data availability, on which a plethora of methods aim to detect and to classify mass abnormalities or tumour cells.

- This work presents two nonlinear system identification models and a multilayer neural network as new choices for digital image processing methods for detecting and classifying breast tissue abnormalities.

- Previous system identification techniques have produced efficient models in input-output dynamical systems. However there is little to say in different domains, for instance, problems involving static (non-time-varying) variables such as feature extraction, classification, pattern recognition or medical image analysis systems.

- The use of nonlinear system identification in image processing confirms its modelling accuracy and widens its application field not only in the medical field but any image processing application related to analysis, detection and classification tasks.

- This chapter outlines the background and theory of image processing, system identification, neural networks and fuzzy logic to shape new machine learning methods for breast cancer detection.

- The capacity of producing transparent image models enables image analysis to regard newly available information like model structure, model parameters and image-based models to be analysed by response signals.

- The multilayer neural-fuzzy network represents an new integrated machine learning framework that takes advantage of type-2 fuzzy sets for tackling uncertainty in digital image processing.

Chapter 3

Image classification using a 2D-NARX model

In the literature, there are efficient parametric models aimed at characterising medical images. However, these techniques are limited to being exclusively linear in the variables. The NARX methodology is designed to contemplate non-linear decision variables and approximate models with an improved scope. This chapter presents a nonlinear parametric framework to characterise digital images and use such information in classification problems. It includes an adaptation process aimed at transferring the information-type from a dynamic to a two-dimensional system, the adoption of the polynomial NARX model into the solution method and the application of the framework as a medical tool for breast cancer diagnosis. The proposed methodology contains the following steps:

- Data transformation from digital images to the input-output system format.
- Estimation of tailor-made mathematical models derived from digital images.
- Extraction of feature values designed for image description, obtained in turn from the image models.
- Classification and detection by using a distance-based classifier.

Tests in a real application in medical images served to evaluate the proposed method. The report of results is in Section 3.3.

3.1 Introduction

Digital image processing and analysis is a growing interdisciplinary branch of signal processing and computer science. Its central aim is to obtain meaningful information from visual patterns, while more advanced human-like skills as classification and recognition rely on supervised, semi-supervised or unsupervised feature analysis [207],[208]. In this context, image classification and pattern recognition techniques make available multiple innovative choices to applied science, with remarkable attention to medicine and bioscience which regularly deal with image analysis of 2D, 3D and time sequences at different zoom levels that rely on precise classification and feature extraction procedures [32].

Within medical analysis, computer-aided detection and diagnosis (CAD) is a significant field of research at present. Although the scope of CAD covers the examination of a variety of medical problems, breast cancer detection is one of the most recurring CAD applications. A pragmatic reason for this is that for instance in the United States, breast cancer is placed at the first place of new cases by cancer type and lies in the second cause of cancer deaths in women by 2017 [23]. The best known therapeutic strategy to reduce breast cancer mortality is early-stage tumour detection, so enhanced CAD algorithms for this goal represent a genuine alternative to save human lives. Parametric system identification methods for image processing have enriched the CAD procedures portfolio by taking advantage of a model-based viewpoint to obtain condensed parameters sets to be used in feature analysis.

The logic behind this detection methods is to attain image models from experimental data, similarly to obtaining parametric time series models for weather forecasting. As time series analysis, parametric CAD system identification methods learn from the image data and help to estimate fundamental values useful for detection processes. Common CAD techniques based on parametric system identification are only capable of identifying linear models, a fact that reduces their flexibility for exploring higher order relationships within images. Note that the term *linear* refers to an expression or model in which none of its variables has been raised beyond the first power. In this regard, the authors of [4] concluded that manoeuvres on digital images are not linear, although linear procedures can approximate these systems in some circumstances. Besides, linear manipulations in digital images may lead to poor results when there is noise with different statistics to Gaussianity [5].

Analogously, the authors of [6] found that linear filters in image processing can miss important image features, such as borders separating background and objects. Bearing this in mind this work adopts the NARX model [88] to improve the characterisation of digital images through the construction of flexible-order image models. Previous work on CAD systems for breast cancer based on parametric system identification models include the AR model [199],[12], the FARMA model [200],[202], the ARIMA model [79] and the ARMA model [10],[11],[203]. Chapter 2 reports these parametric methods in detail.

3.2 The 2D-NARX methodology

The recommended image processing method, termed 2D-NARX, can be divided into four modules: image rendering, image modelling, feature vector extraction

and classification, with all parts coded to run in sequence and automatically. The proposed CAD scheme derived from 2D-NARX seeks to detect suspicious patterns by adding initial and final divide-and-conquer steps, as used in [10],[209] and documented earlier for image processing in [210]. The integrated 2D-NARX approach is summarised in 3.1.

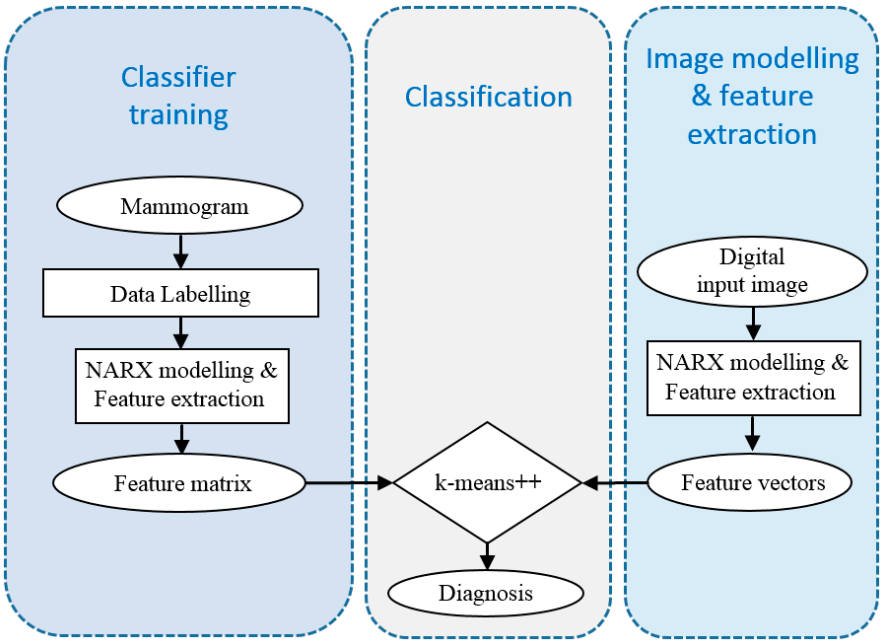


Figure 3.1: 2D-NARX general algorithm chart.

The work flow of the classification and detection methodology includes a separate training process made in advance the single image feature extraction. For this latter stage, the processing starts with the initial partitioning of the input image (divide-and-conquer algorithm), followed by the transformation of the image data to input-output data, the NARX image modelling, the FROLS model structure detection, the feature extraction and the classification and diagnosis via the K-means algorithm. The following sections detail each one of this steps.

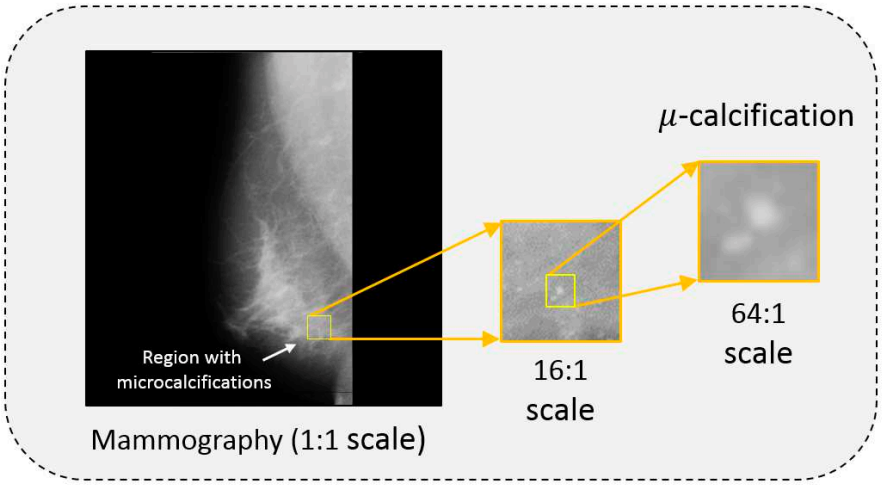


Figure 3.2: Magnification of microcalcifications in mammogram mdb233 (from [212]).

3.2.1 Digital mammogram partitioning

The problem of breast cancer detection in digital images poses as a central element the tumour characterisation, which have a considerably smaller size than the whole digitised mammogram film. For this reason, a divide-and-conquer strategy was adopted to obtain useful dimensioned subimages out of the initial mammogram, zooming in on the tissue abnormalities including mammary microcalcifications, which represent a critical element in breast cancer detection [211]. In other words, the divide-and-conquer strategy aims to solve a massive problem by breaking it down into smaller problems and use the solutions of each to solve the original instance [210]. Please note that the medical image processing literature terms the area enclosing the object of interest as the *region of interest* (ROI).

To illustrate the need for the image partitioning in the breast classification problem, Figure 3.2 seeks to exemplify the significant difference in proportions between a complete mammogram and a microcalcification in an image from the

Table 3.1: Average tumour diameter of the MIAS database [212].

Tumour class	**Average size (pixels)**
Benign	43.39
Malign	53.56
All classes	49.56

selected database. In the example, it is possible to observe with the naked eye that a classification analysis at full mammogram level may easily leave microcalcifications out of scope, whose failure to be detected is dangerous in medical terms since, according to [211] these are catalogued as *extremely suspicious* in 78% of cases and might well represent the only noticeable trace of a tumour.

The divide-and-conquer strategy includes the concept of *subimage*, which in size terms is analogous to the ROI. The subimage size selection is a crucial decision element since the bigger it is, the more data is available to find a reliable image model, where the data represent the intensity value of the pixels. However, if the subimage size is more extensive than necessary, pixels from other classes may hinder the classification process [10]. To estimate the best suitable choice according to the breast cancer detection problem, the average tumour diameter by class from the selected database [212] was calculated to make a more reality-based decision (Table 3.1).

The selection of a 64×64 pixels subimage size was considered suitable to enclose the ROIs of the MIAS database effectively. This number also matches to a large extent the partition of the original mammogram in a 1/16 ratio per side recommended in [10],[11]. This proportion takes into account that the database

mammograms present 1024 pixels by side so that dividing 1024 by 16 the result is 64 pixels per subimage side. Therefore, the calculation of the number of subimages by mammogram is as follows:

$$s = \frac{HW}{N^2} \tag{3.1}$$

where H and W are the height and width of the original image respectively and N the sub-image size per side, giving 256 subimages for the problem faced in this chapter.

3.2.2 Two-dimensional image rendering and representation

The modelling process is in system engineering as a bridge connecting a real system with a solution algorithm. Analogously, the image interpretation from the input-output systems viewpoint is a vital step of the methods proposed in this and the following chapter. Given that the NARX models are designed by nature to represent input-output systems, special attention had to be given during the interpretation process to address the goal of adapting the two-dimensional image field into the NARX paradigm consistently.

Figure 3.3: Multiple-input and single-output system.

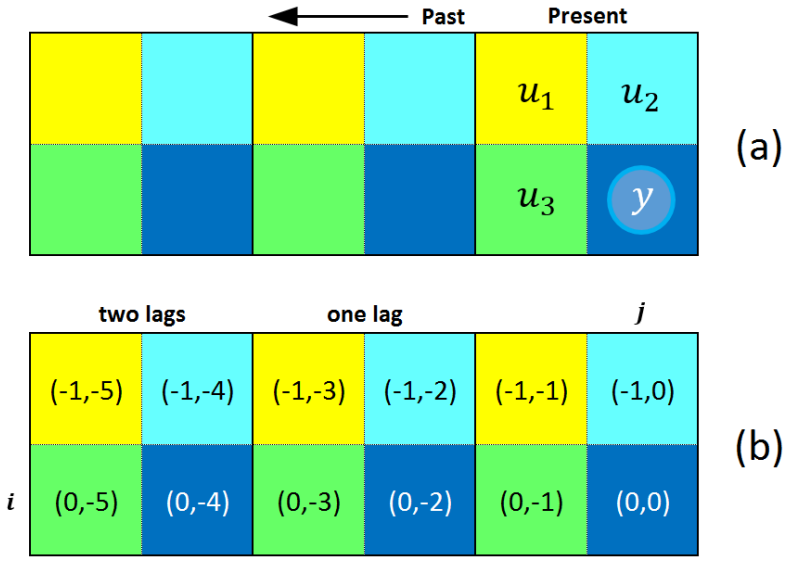


Figure 3.4: MISO system modelling of the image field.

To attain that objective, a pixel neighbourhood which could allocate a single output ($y(k)$) and three adjacent exogenous inputs ($u_1(k)$, $u_2(k)$ and $u_3(k)$) was chosen for the sake of a straightforward but sufficient modelling in terms representativeness. Figure 3.3 shows a diagrammatic example of a multiple-input, single-output (MISO) system which is unknown.

The final configuration of four variables (one output and three inputs) is analogous to a 2×2 bidimensional symmetric neighbourhood of adjacent pixels, obtaining in this way an equivalent to a MISO system projected in two dimensions.

Figure 3.4 illustrates the resulting data conversion. Scheme (a) shows the position of the input-output variables within the pixel neighbourhoods, where the more on the left the neighbourhood is located, the more lagged it is from the time series point of view. Scheme (b) represents the same pixel block as (a) but instead of displaying the variable names, it shows the corresponding coordinates (i, j). The above is achieved by fixing the origin $(0, 0)$ at pixel y, which is equivalent to the

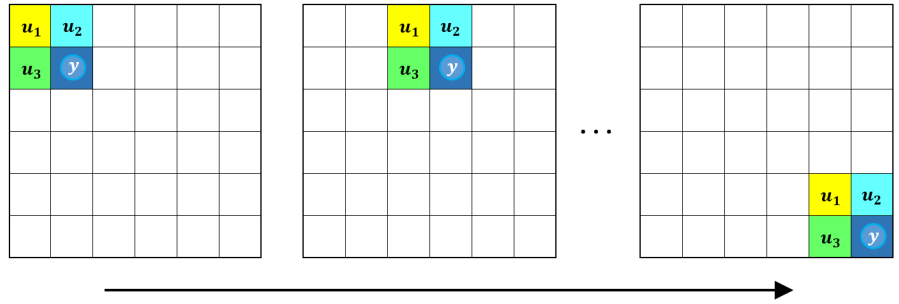


Figure 3.5: Neighbourhood mask scanning movement during the image data collection.

MISO system output. With that reference, the rest of the pixels take coordinate values to be useful in model equations 3.2 to 3.5.

This format allows to consider the image as a series of *observations* (data blocks spread vertically in space) distributed from top to bottom and from left to right, where the lagged observations point towards the left side direction. In addition to the above, no overlaps between the present and past observations are allowed to avoid getting naive or simplistic models (for instance, a model consisting of only the $y(t-1)$ variable).

The Figure 3.5 depicts the raster scan direction for collecting the intensity map of an image from a dynamical system perspective. The following section describes the mathematical formulation corresponding to this structure according to the NARX model.

3.2.3 The NARX model

As seen in the previous chapter, system identification procedures aim at formulating systems models from observational data without prior information of the model structure. Several of these systems can be challenging because of their in-

tricacy, may include non-Gaussian variables or might exhibit rapid changes from one period to another, such that nonlinear approaches stand out from the rest as good modelling and approximation choices.

However, the nonlinear condition conveys higher computational costs and modelling accuracy challenges for system identification models. To overcome these difficulties, the nonlinear autoregressive moving average with exogenous input (NARMAX) model [110] (Section 2.2.1) was presented as a valid choice to model a wide range of dynamic problems by processing past inputs and outputs to explain the system output [88]. The nonlinear autoregressive with exogenous input (NARX) model is a simpler version of the NARMAX model that omits previous errors, in the form of regressors, during the identification task for the sake of a simpler representation and a greater ease to be solved at the cost of a minor precision loss [213],[113].

Therefore, the choice of one of the above options depends on the priority for simplicity and ease of solution versus accuracy [214]. The present study deals with a problem that requires the rapid processing of a large number of subimages. Also, preliminary tests showed that the regressors of a MISO system where sufficient to describe both reduced and medium size images, so the NARX model was chosen as it offered the right balance. The general NARX mapping for a single-input and single-output system (SISO) was described earlier in Equation 2.2.

The NARX mapping proposed for image processing, derived from the formula (2.2) and the mathematical 2D modelling, as detailed in Figure 3.4 and Section 3.2.2, is as follows (note that the related description is found after the formula).

$$
\begin{aligned}
y(i,j) = F[&y(i,j-2(1)), y(i,j-2(2)), ..., y(i,j-2n),\\
&u_1(i-1,j-1-2(1)), u_1(i-1,j-1-2(2)), ..., u_1(i-1,j-1-2n),\\
&u_2(i-1,j-2(1)), u_2(i-1,j-2(2)), ..., u_2(i-1,j-2n),\\
&u_3(i,j-1-2(1)), u_3(i,j-1-2(2)), ..., u_3(i,j-1-2n)] + e(i,j)
\end{aligned}
\tag{3.2}
$$

where $y(i,j)$ is the system output, $u_1(i-1,j-1)$,$u_2(i-1,j)$,$u_3(i,j-1)$ are the system inputs, $e(i,j)$ is an independent noise sequence, $n = n_u = n_y$ are the maximum lags defined for the system variables (maximum observations in the past contributing to explain the system output) and $F[\cdot]$ is a nonlinear function to be defined. The regressors in Equation 3.2 represent space instead of time, unlike common NARX formulations for dynamic systems. Indices i, j represent the coordinates for rows and columns, respectively distributed in the pixel mesh.

By following Figure 3.4, the squares coloured in navy blue show the y variable at point $(0,0)$ and the corresponding lags starting to the left at point $(0,-2)$. Then, these go on to point $(0,4)$ and so on. The first line of Equation 3.2 contains that sequence. Similarly, coloured in yellow, the u_1 variable at $(-1,-1)$ has lagged versions to the left from point $(-1,-3)$, then at $(-1,-5)$ and so on. The second line of Equation 3.2 contains that sequence. The sequences for the variables u_2 and u_3 can be inferred in the same way by following the green and light blue squares succession, respectively.

The polynomial NARX

At this point is it equally relevant to note that the NARX equations 2.2 and 3.2 describe only a relationship between the system output and the unknown nonlinear

function which depends on several lagged variables (listed in brackets within the function) plus an independent noise sequence. Such a situation makes it necessary to choose a model/blueprint from the broad range of choices (see Section 2.2.1 from the previous chapter) capable of expanding the NARX mapping function stated for image processing in Equation 3.2. From all alternatives, the polynomial NARX models are the most popular given the advantages to be discussed later on [88].

A way of seeing the polynomial formulation is by considering that it equals the output of a process to several monomials or power products representing all the possible combinations of predecessors variables contained in the non-linear function $F[\cdot]$ of the NARX mapping of Equation 3.2, from degree 1 (linear) up to the nonlinear degree l. The power-form polynomial NARX models offer the following advantages:

- The models are transparent, legible and easy to write.
- Given that these models are smooth functions, solutions are more precise and run times are reduced.
- Flexibility to describe a wide variety of nonlinear systems in the time domain, the frequency domain or a combination of both.
- Availability of widely tested algorithms to efficiently solve the formulation.

The following equation rewrites Equation 3.2 from the explained above 2D NARX mapping and the general definition of the power-form polynomial NARX (Equation 2.3):

$$\begin{aligned} y(i,j) = \theta_0 + \sum_{i_1=1}^{n} \theta_{i_1} \mathrm{x}_{i_1}(i,j) + \sum_{i_1=1}^{n}\sum_{i_2=1}^{n} \theta_{i_1 i_2} \mathrm{x}_{i_1}(i,j)\mathrm{x}_{i_2}(i,j) + ... \\ \sum_{i_1=1}^{n} ... \sum_{i_\ell=i_{\ell-1}}^{n} \theta_{i_1 i_2 ... i_\ell} \mathrm{x}_{i_1}(k)\mathrm{x}_{i_2}(i,j)...\mathrm{x}_{i_\ell}(i,j) + e(i,j) \end{aligned} \tag{3.3}$$

where l is the maximum nonlinear degree, $\theta_{i_1,i_2,\cdots,i_m}$ are the model parameters, $n = n_y + n_{u1} + n_{u2} + n_{u3}$, where $n_y, n_{u1}, n_{u2}, n_{u3}$ are the maximum lags for the system outputs and inputs respectively, and $e(i,j)$ is an independent noise sequence. Taking into account the general description (Equation 2.4) the vector of basic cross-coupled regressors $x_m(i,j)$ combined in the 2D polynomial expansion (Equation 3.3) is defined for a MISO system as:

$$x_m(i,j) = \begin{cases} y(i, j-2m), & 1 \le m \le n_y \\ u_1(i-1, j-1-2(m-n_y)), & n_y+1 \le m \le n_y + n_{u_1} \\ u_2(i-1, j-2(m-n_y-n_{u_1})), & n_y + n_{u_1} + 1 \le m \le n_y \\ & +n_{u_1} + n_{u_2} \\ u_3(i, j-1-2(m-n_y-n_{u_1}-n_{u_2})), & n_y + n_{u_1} + n_{u_2} + 1 \le m \le \\ & n_y + n_{u_1} + n_{u_2} + n_{u_3} \end{cases} \tag{3.4}$$

The visualisation of the regressor sequence $x_m(i,j)$, recently described in Equation 3.4, was simplified by fixing the maximum lags as the model was indeed

configured in this work (see Section 3.3), so that $n_y = n_{u1} = n_{u2} = n_{u3} = 1$. In this way, the next equation expresses the regressor sequence in (3.4) as follows.

$$x_m(i,j) = \begin{cases} y(i, j-2m), & 1 \leq m \leq 1 \\ u_1(i-1, j-1-2(m-1)), & 2 \leq m \leq 2 \\ u_2(i-1, j-2(m-2)), & 3 \leq m \leq 3 \\ u_3(i, j-1-2(m-3)), & 4 \leq m \leq 4 \end{cases} \tag{3.5}$$

With the adaptation of the NARX model to a two-dimensional viewpoint and the definition of the nonlinear function expansion, adapted as well to the 2D problem, the remaining step for system identification is to solve the power-form polynomial model. Below is the explanation of the method adopted for this end.

3.2.4 FROLS model structure detection

The NARX system identification philosophy bases itself in the first place on the expansion of a (previously unknown) nonlinear function followed by the construction of an extensive dictionary D containing M elements or terms. The components of the dictionary, known as well as candidates, are obtained from the function expansion, and it is from this dictionary that the structure detection algorithm selects the final model terms. When the NARX function follows the power-form polynomial expansion, it's important to emphasise that each candidate represents a power product (model term) of the polynomial formulation, taking into account that each of these monomials is composed of cross-coupled system input and output variables.

Figure 3.6 depicts the processing flow from data representation to the nonlinear function expansion, a necessary step preceding the model structure detection.

After the polynomial expansion, the forward regression orthogonal least squares (FROLS) algorithm [183] (see Section 2.3.1) was incorporated into the algorithm given its effectiveness to make one-at-a-time orthogonalised steps intended at selecting the best available candidate terms contained in the polynomial function. The incorporation took place by unifying the orthogonal least squares (OLS) [185] and the Gram-Schmidt algorithms [187].

The OLS algorithm helps to select the candidate terms with higher error reduction ratio (ERR) via the Gram-Schmidt orthogonalisation algorithm, which promotes the stepwise selection of unselected candidate terms adding complementary information to that of the model terms already included in the solution, generating parsimonious and accurate models. Figure 3.7 summarises such structure selection process, where the resulting model is made up of the sum of power functions (terms) and model parameters.

The described process, from the 2D/image processing perspective, initially reads, transforms and represents the input image according to the polynomial

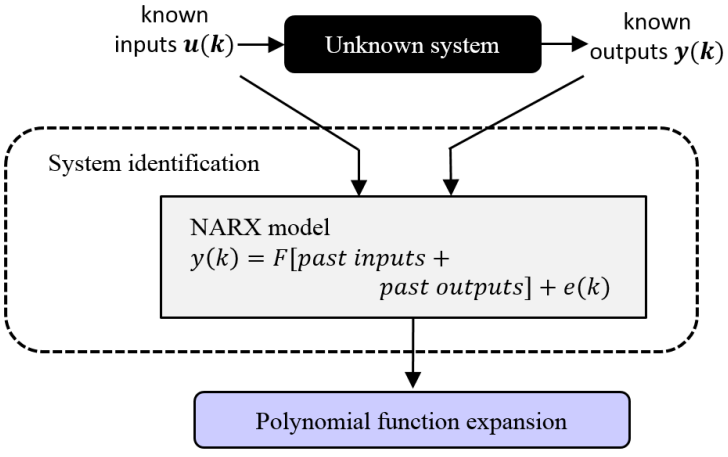


Figure 3.6: Flowchart of the NARX mapping and polynomial expansion of the nonlinear function.

NARX model in a data matrix or dictionary D. Afterwards, this dictionary is processed by the FROLS algorithm, which creates a realistic and parsimonious NARX model of the input image. The image model is taken hereinafter to obtain representative feature values useful in classification and detection algorithms.

3.2.5 Extraction of feature values

Once the system identification design was complete, a significant challenge related to the construction of feature values resulting from the ROI image models was to be solved. In spite the NARX-FROLS system modelling offers competitive advantages to accurately identify tailor-made models for various real-life systems regardless of whether these are linear or not, such adaptive modelling advantage brings along the difficulty of producing equal-sized feature vectors (necessary in the classification process) out of different models with variable/adaptive NARX structure from different ROIs. Therefore, a fixed number of unchanging input signals were carefully selected to solve such hindrance. These input signals played the role of stimuli of the model's behaviour.

Such behaviour took the form of output signals from which it was possible to extract information. The input signals selection aimed at producing mutually

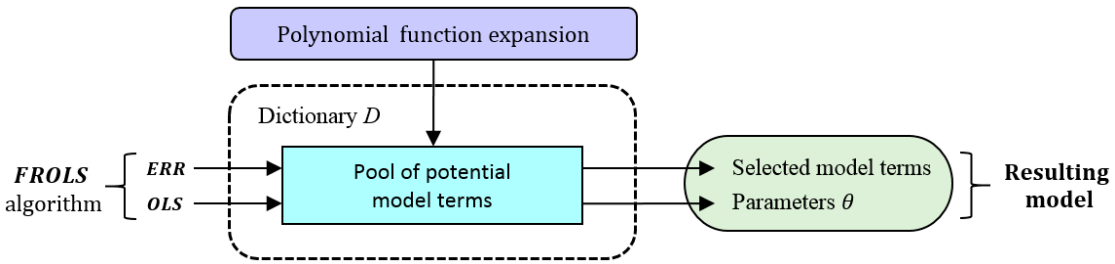


Figure 3.7: FROLS's model selection flowchart.

uncorrelated model responses, so a series of tests were made by taking into account diverse instances.

Such stimulus-response design was built up by taking into account the NARX image model as a black box and a set of fixed input signals to produce the same number of response (output) signals. The diagram of the model stimulation procedure is detailed in Figure 3.8.

After that, the approximation of feature vectors of the same length (required for classification) resulted from using four statistical estimators on the response signals obtained from the stimulus-response design. The measure selection aimed at obtaining a statistical description or featuring of the response signals as reliable as possible. The selected estimators were the *population mean* (μ), the *standard deviation* (SD), the *interquartile range* (IQR) and the *mean absolute deviation* (MAD). In statistical terms, the first one represents a central tendency while last three are dispersion measures.

More profoundly, the population mean, which is a measure considering all data values, is regarded as a reliable and representative estimator, while the standard deviation is an accurate estimator indicating the spreading out level of the data

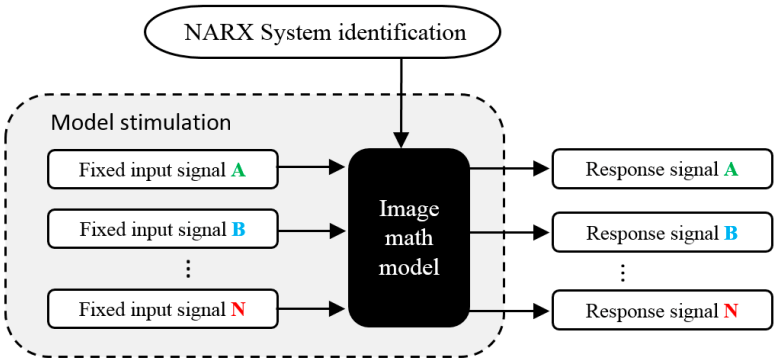


Figure 3.8: Stimulating the model's behaviour.

points concerning the mean, being suitable for non-chaotic distributions with minimum outlier presence.

The interquartile range and the mean absolute deviation, represent robust-type measures, ideal for characterising mixed or heavy-tailed data distributions. The IQR is the difference between the third and the first quartile. The quartiles are three population values dividing the data set into four equal parts, so the IQR is practically unaffected by extreme values. The MAD is the average of the data points distances regarding the median value and provides a clear idea of the data set variability while extreme values hardly distort it. The formulas of the estimators are listed below along with their main descriptive advantage.

$$\textit{Population mean } (\mu) \qquad \frac{\sum_{i=1}^{N} x_i}{N} \tag{3.6}$$

$$\textit{Standard deviation } (SD) \qquad \sqrt{\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}} \tag{3.7}$$

$$\textit{Interquartile range } (IQR) \qquad \textit{Quartile } 3 - \textit{Quartile } 1 \tag{3.8}$$

$$\textit{Mean absolute deviation } (MAD) \qquad \frac{\sum_{i=1}^{N} |x_i-\mu|}{N} \tag{3.9}$$

The last characterisation step poses the mentioned statistical estimators to analyse the output signals of the model to get a fixed number of feature values, which are concatenated at the end of the procedure to build a feature vector. Figure 3.9 shows the flow of this process. Also, it was decided to include in the general procedure an extra position of the input image in the form of a 90-degree rotation to increase the algorithm's ability to recognise spatial features in the ROI image, such as tumour position, microcalcifications and other abnormalities.

With the previous design, the feature vector length representing the input image is equal to the product of three values: statistical measures (4), stimulation signals (6) and rotation angles (2), producing in total $4 \times 6 \times 2 = 48$ feature values.

3.2.6 Classification and detection

The classification and diagnosis process of the method presented in this chapter bases its power on the well-known K-means algorithm [78]. The K-means (see Section 2.1.5) is a distance-based iterative refinement method (also considered a machine learning method) aimed to divide a set of input observations into a k number of groups. This procedure is done by (1) choosing at random K centres, (2) linking the observations to their closest centre and (3) reallocating the centres through the arithmetic mean of the cluster, (4) repeat steps 2 and 3 until the system converges to a stable state.

Unlike the traditional K-means clustering algorithm, which applies unsupervised learning to processing unlabelled data, the algorithm used in this work was instead supervised. This process was done by labelling the known data in the first

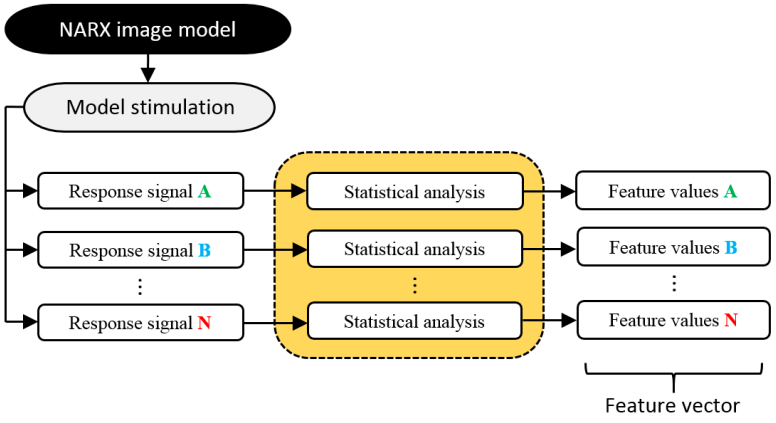


Figure 3.9: Feature extraction by stimulating the model behaviour.

place, taking into account the database documentation regarding the tumours coordinates, diameters and classes (benign or malign), as well as other relevant visual features contained in the mammograms, including tissue density (fatty, glandular or dense) and background artefacts as tags, pectoral muscle and imaging and scanning mistakes.

Then the labelled data are taken as patterns or first centres on which the algorithm makes a series of linkage-directed iterations between the latest and the input image feature vectors. Unlike the original, this modification allows the classifying process to associate in the multidimensional space, via the shortest Euclidean distance, the input image with the pre-labelled training images.

A second modification considered here is the adoption of an enhanced K-means algorithm, named by its authors the K-means++ [215]. The algorithm instead of merely assigning the input vectors to the nearest centre, weights the first ones according to their square distance to the cluster centres, producing by this simple modification a much faster convergence and a more accurate classification compared to the original algorithm [215]. The proposed breast cancer detection method in digital mammograms compares the input image feature vector concerning the human labelled feature matrix, which is taken as a benchmark by the K-means algorithm to determine the input image medical condition. The overall classification process design is charted in Figure 3.10.

To explain more in detail the classification itself, the K-means++ partitioning takes the new, unlabelled input vectors to infer if these are whether healthy or positive (suspicious) swiftly. A message is displayed by the program to the user whenever a suspicious case is found, along with the inferred abnormality class (whether it is benign or malign). The diagnosis processing design of a full mammogram considers that it is enough for the classifier to identify at least one single

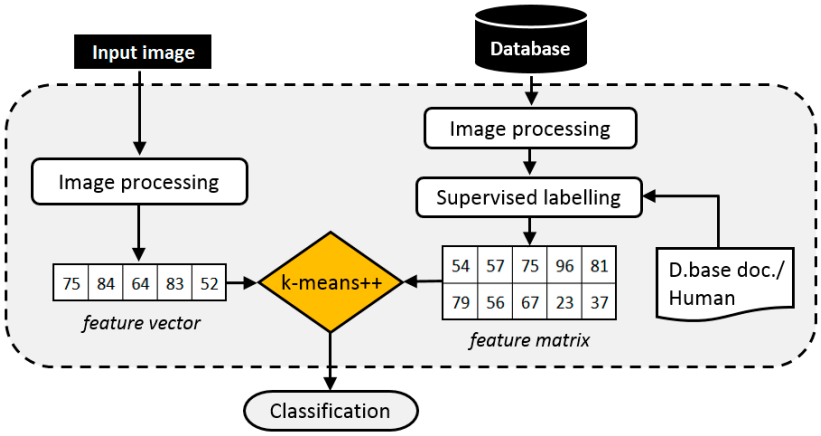


Figure 3.10: Flow diagram of the classification and detection design.

vector as suspicious, out of the 256 composing the X-ray mammography, so that the whole mammography be considered a positive case that needs attention and follow-up by specialised human medical personnel.

3.3 Experiments and results

The experimental section of the CAD-NARX method intends to describe in detail the selected case study, the challenges faced during the labelling of the samples, the resulting image models and the obtaining of the performance values of the classifier. As mentioned earlier, the development of the proposed method focuses on extracting a compacted series of feature values capable of representing a digital image of any kind as reliable and fast as possible to produce data useful to classification. After that, the breast cancer detection problem, a significant public health problem, was selected to appraise the image processing method.

3.3.1 Case of study

The specific case of study for the breast cancer detection problem was the mini-MIAS database of mammograms [212]. This publicly available image set resulted from a selection of digitised X-ray films obtained from a single primary health centre associated with the United Kingdom National Breast Screening Programme. This free-access repository contains 322 films, scanned and digitised at 50-micron pixel edge, reduced to 200-micron pixel edge so that images are 1024 x 1024 pixels. Besides, the dataset includes detailed documentation on the character of the background tissue and, where applicable, class, severity, coordinates and approximate radius of abnormalities.

All mammograms present a medio-lateral oblique (MLO) view and are greyscaled. A greyscale image is a single-channel digital representation which, unlike 3-channelled RGB images, pixel values symbolise only a quantity of light (an intensity value ranging from 0 to 255) so that several shades of grey may equal or lay in between these values.

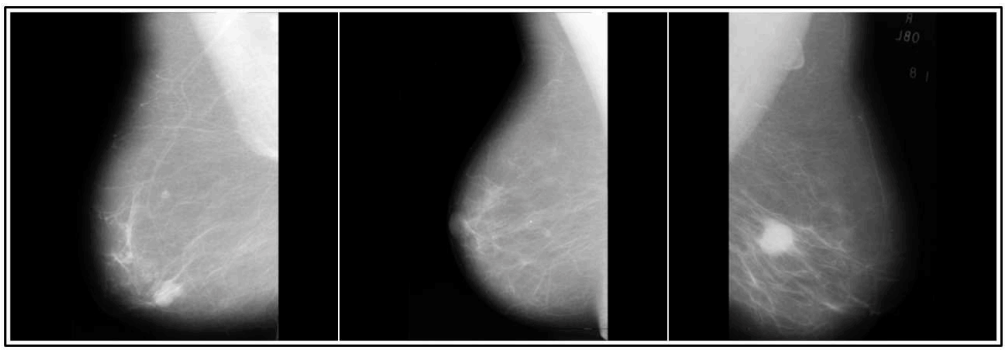


Figure 3.11: Typical medio-lateral oblique view of images mdb005 -benign-, mdb009 -healthy- and mdb028 -malign- (from [212]).

Table 3.2: Database mammogram class distribution [212].

	Benign	**Malign**	**Normal**	**Total**
Count	66	52	204	322
Percentage	20.50	16.15	63.35	100

Figure 3.11 shows examples of three breast-type specimens in the selected database. In the image on the left side there is an X-ray scan of a patient with a benign anomaly, in the centre a mammogram of a healthy patient and to the right a plaque of a patient with a malignant tumour.

Note that for the convenience of the visualisation the X-ray films always present the image background in black, so that the presence of any mass or object appears in shades of grey becoming clearer in direct proportion to the level of obstruction of these to the passage of the dark background.

In that way, the denser is an object or abnormality appearing in the mammography area, the thicker and whiter it looks. The average tumour radio of the image set is 43 pixels for benign and 53 pixels for malign growths. 33% of images correspond to fatty, 32% to fatty-glandular and 35% to dense breast type. Table 3.2 displays a summary of the database categories distribution.

3.3.2 Setting up of the model parameters

The experimentation process includes tumour detection tests aimed at knowing the efficiency of the feature extraction method to discriminate between images with and without tumour occurrence. Separately, experiments were conducted to see the discrimination accuracy between benign and malign tumours. A pondered random

sampling was performed within each class to select the training and testing samples to preserve the overall class distribution of healthy, benign and malign images of the mini-MIAS data set. The original image set was randomly split and into 222 mammograms for training and 100 mammograms for testing. Such a partition (69% training and 31% testing) was selected to maximise the number of training images and to leave at the same time a large enough number of ROI images for testing.

The first congruency tests of the new classifier utilised small and straightforward image sets, including symbol and letter libraries. These tests sought to verify that the classifier was capable of sorting in an unsupervised way the letter and symbol images with mutual similarity within the same groups.

After verifying the success in this first step, the 2D-NARX identification algorithm came into play in the processing of real mammograms. It was adjusted to a maximum nonlinear degree $\ell = 2$ and maximum lags n_y, n_{u1}, n_{u2}, $n_{u3} = 1$, as it was the best balance found regarding representativeness and simplicity. The ROI size was equal to 64×64 pixels given the average tumour size of the database. The error tolerance of the ERR stop-criterion was 0.15%. The programs were coded and run in MATLAB R2014b in a computer Dell Optiplex 7020 with IntelCore i5-4590 CPU at 3.30GHz.

3.3.3 Data labelling and supervised learning

In spite of the high-quality database, numerous artefacts and scanning imperfections within images were present, such as unknown breast position and orientation (left or right), duct tapes, orientation tags, low-intensity labels and scanning artefacts, as described in [216] for the same database. Besides, there were uneven

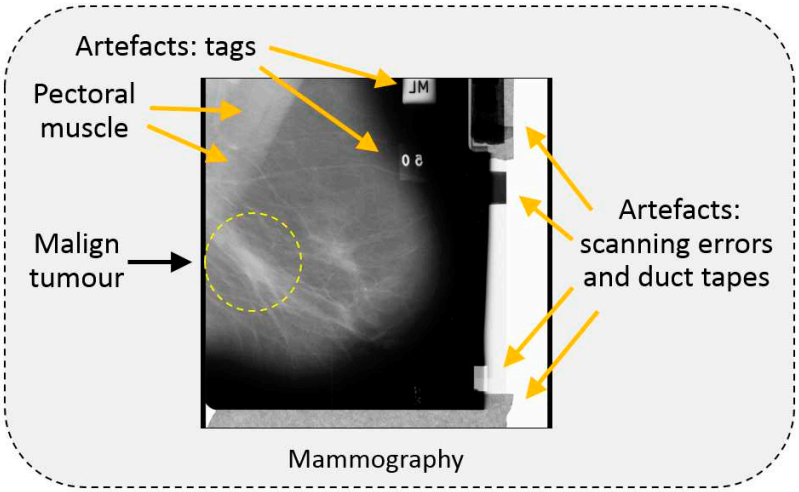


Figure 3.12: Example of a mammogram with several background artefacts (image 274mdb [212]).

contrast levels as well as breast tissue and pectoral muscles dissimilarly positioned along the images. An example of such difficulties is provided in Figure 3.12, where the presence of various artefacts represents a potential hinder for the classification process.

The above difficulties led to carry out a careful data labelling given that the cancer detection problem involves the care of human lives. The process derived 5335 feature vectors of healthy, benign and malign digital ROI images. One by one, the mammograms of the test set were subject to comparison with the training data via the K-means classifier. To decide whether an X-ray mammogram is suspicious, at least one of its 256 subimages must fall into the benign or malign category. Otherwise, the entire image is tagged as healthy (see details in Section 3.2.5).

3.3.4 Resulting image models

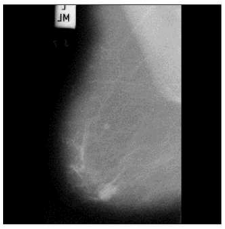
This section intends to provide examples of 2D-NARX models identified by the FROLS algorithm, representing three clinical cases in mammograms containing a benign, a healthy and a malign sample, according to the documentation attached in the database. The sample selection took into account the uniformity and background clarity to ease the visual examination, especially with regard to the presence of tissue abnormalities. The selected images correspond to those in Figure 3.11 (Section 3.3.1, Case of study) and their full view is in Appendix A.

The information contained in tables 3.3 and 3.5 to 3.9 includes, from left to right, the thumbnail of the ROI, the index of the model terms sorted out according to their ERR, the polynomial term included in the NARX model as a sum, the parameter θ_i linked to the model term, and finally the ERR estimator, which indicates the contribution in percentage of the term to decrease the model prediction error concerning the observed data. Note that Equation 3.10 shows how the model expressed in Table 3.3 takes the form of a mathematical expression.

Models from full mammograms

The first tables correspond to NARX models obtained directly from mammograms, that is, from 1024×1024 pixel images of the case study with no close-up or zoom added over any ROI such as tumours or microcalcifications. The aim of studying this group is to highlight the proposed method capacity to generate mathematical models with terms, parameters and structures tailored to the input images. Also, the exercise seeks to point out the difficulty involved in characterising any abnormal samples without an adequate zooming-in level over the ROIs. Tables 3.3,3.5 and 3.6 present the first set of examples of image models.

Table 3.3: Model of a benign mammogram (1024 × 1024 px, mdb005 [212]).

Mammogram	Index term	Model term	Parameter	ERR
	1	$u_3(k-1)$	0.9930	99.9441%
	2	$u_2(k-1)$	0.8041	0.0106%
	3	$u_1(k-1)$	-0.7950	0.0332%
	4	$y(k-1)u_1(k-1)$	-0.0032	0.0000%
	5	$y(k-1)u_2(k-1)$	0.0028	0.0011%
	6	$u_1(k-1)u_2(k-1)$	0.0038	0.0003%
	7	$[u_2(k-1)]^2$	-0.0033	0.0008%

To better explain the NARX model in 2D, the first step is to write the data in Table 3.3 as an identified model equation as follows.

$$\begin{aligned} y(k) = {} & 0.993u_3(k-1) + 0.804u_2(k-1) - 0.795u_1(k-1) \\ & -0.0032y(k-1)u_1(k-1) + 0.0028y(k-1)u_2(k-1) \\ & +0.0038u_1(k-1)u_2(k-1) - 0.0033[u_2(k-1)]^2 + e(k) \end{aligned} \tag{3.10}$$

Then, Figure 3.13 shows below the basic model relationship of Equation 3.10 with the 2D image distribution. The unknown variable y in the period k is determined by the combination of intensity values of the squares (pixels) u_1, u_2, u_3 and y of the period $k-1$. In other words, the NARX model predicts the value of the y box with the information of other specific boxes located on its left.

By observing Figure 3.13, we can see in Table 3.4 that regressors in the time domain have a two-dimensional equivalence in the new framework.

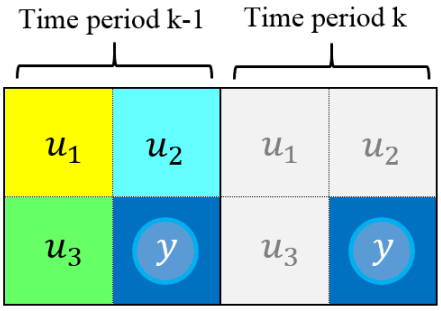


Figure 3.13: The output y in time period k can be explained by combining regressors $u_1(k-1), u_2(k-1), u_3(k-1)$ and $y(k-1)$ in the proposed model. When in the coordinate system y equals to $y(i,j)$, the lagged variables equal to $u_1(i-1,j-3), u_2(i-1,j-2), u_3(i,j-3)$ and $y(i,j-2)$.

Table 3.4: Examples of the equivalence between representation systems for the proposed method.

Representation	Index	Lagged exogenous inputs			Lagged output
Time domain	k	$u_1(k-1)$	$u_2(k-1)$	$u_3(k-1)$	$y(k-1)$
Spatial domain	i,j	$u_1(i-1,j-3)$	$u_2(i-1,j-2)$	$u_3(i,j-3)$	$y(i,j-2)$

Thereby, Equation 3.10 can also be expressed in two dimensions as Equation 3.11 shows next.

$$\begin{aligned} y(i,j) = {} & 0.993u_3(i,j-3) + 0.804u_2(i-1,j-2) - 0.795u_1(i-1,j-3) \\ & -0.0032y(i,j-2)u_1(i-1,j-3) + 0.0028y(i,j-2)u_2(i-1,j-2) \\ & +0.0038u_1(i-1,j-3)u_2(i-1,j-2) - 0.0033[u_2(i-1,j-2)]^2 + e(i,j) \end{aligned} \tag{3.11}$$

The first example in this section (Table 3.3) displays the NARX model of a mammogram which presents two benign tumours practically spliced in the lower zone of the breast. Although the method was able to model the whole digitised film accurately (as the sum of ERR precision values indicates), it was problematic for the new framework (and presumably for any other analysis) to lead to an accurate classification without adequately zooming in over the ROI.

Table 3.5 and Table 3.6 represent the models of mammograms with healthy and malignant cases respectively. It is possible to see that the first three models are broadly similar to each other in terms of the presence of regressors made up of the lagged exogenous inputs $u_1(k-1), u_2(k-1)$ and $u_3(k-1)$. As expected, the no-zoom condition entails a comprehensive characterisation of the total image at the expense of lower reliability in the classification of tumours, which are nearby 1/256 smaller in proportion.

Table 3.5: Model of a healthy mammogram (1024 × 1024 px, mdb009 [212]).

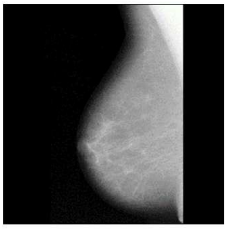
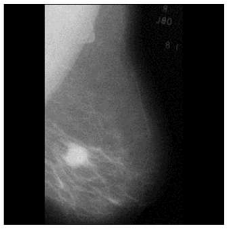
Mammogram	Index term	Model term	Parameter	ERR
	1	$u_3(k-1)$	1.0944	99.9021%
	2	$u_2(k-1)$	0.3031	0.0243%
	3	$u_1(k-1)$	-0.3907	0.0382%
	4	$u_1(k-1)u_3(k-1)$	-0.0026	0.0002%
	5	$u_2(k-1)u_3(k-1)$	0.0026	0.0073%
	6	$[u_3(k-1)]^2$	-0.0004	0.0002%
	7	$y(k-1)u_1(k-1)$	0.0004	0.0001%

Table 3.6: Model of a malign mammogram (1024 × 1024 px, mdb028 [212]).

Mammogram	Index term	Model term	Parameter	ERR
	1	$u_2(k-1)$	0.9617	99.9605%
	2	$u_3(k-1)$	0.2648	0.0012%
	3	$u_1(k-1)$	-0.1232	0.0047%
	4	$y(k-1)$	-0.1029	0.0004%
	5	$y(k-1)u_2(k-1)$	0.0008	0.0000%
	6	$[u_2(k-1)]^2$	-0.0029	0.0049%
	7	$u_2(k-1)u_3(k-1)$	0.0021	0.0004%

MODELS FROM MAMMOGRAM PARTITIONS

The second group of models represent subimages of 64 × 64 pixels obtained from the mammogram partition, according to the criteria specified in Section 3.2.1, which details the *divide and conquer* procedure used in this chapter.

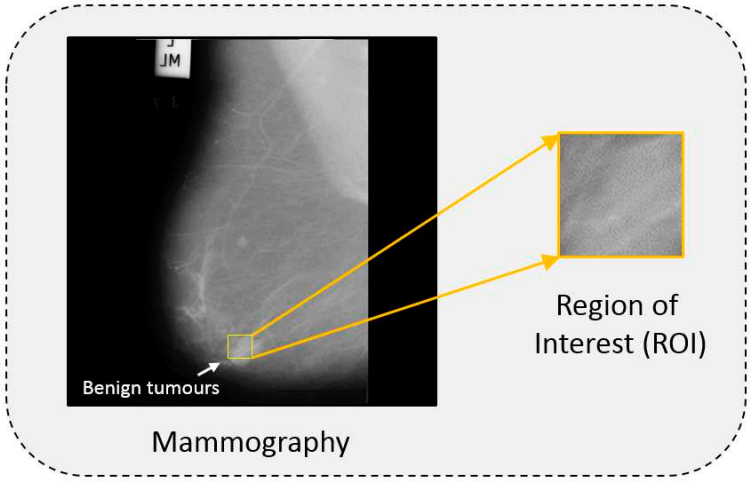


Figure 3.14: ROI from a malign sample (mammogram mdb005 [212]).

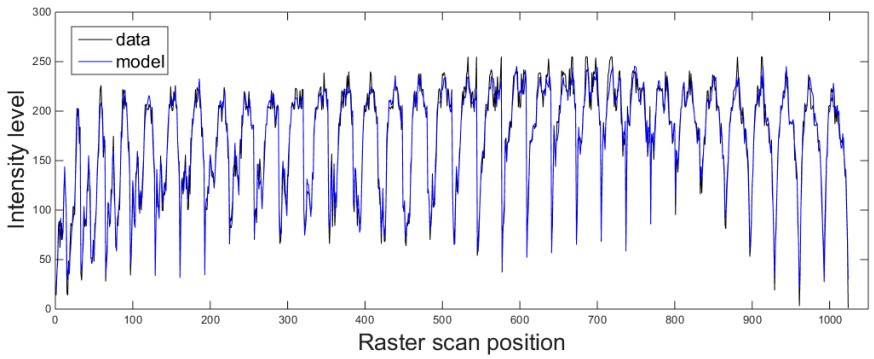


Figure 3.15: Model fitting to data, benign ROI from mdb005 [212].

Figure 3.14 shows the proportional difference between a subimage and a complete mammogram. Note that that the tumour region is better circumscribed by the ROI area compared to the full image. The full visualisation or zero zoom level includes areas of marginal interest for the study.

Three tables showing the NARX subimage models are presented below along with comparative graphs displaying the model predicted output ($\hat{y}(k)$) and the real observed output ($y(k)$) to appraise the model fit accuracy. Tables 3.7 to 3.9 contain the subimage models and Figures 3.15 to 3.17 display the the model fitting graphs.

Table 3.7: Model of a benign ROI (64 × 64 px, mdb005-217 [212]).

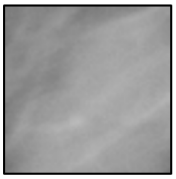
Subimage	Index	Model term	Parameter	ERR
	1	$u_3(k-1)$	0.8177	99.7130%
	2	$u_2(k-1)$	0.7274	0.0546%
	3	$u_1(k-1)$	-0.6003	0.1116%
	4	θ_0	6.1695	0.0048%
	5	$[u_3(k-1)]^2$	0.0011	0.0007%
	6	$y(k-1)u_3(k-1)$	-0.0013	0.0027%
	7	$[y(k-1)]^2$	0.0003	0.0016%

Table 3.8: Model of a healthy ROI (64 × 64 px, mdb009-184 [212]).

Subimage	Index term	Model term	Parameter	ERR
	1	$y(k-1)$	0.7025	98.0504%
	2	θ_0	17.8002	0.0977%
	3	$y(k-1)u_3(k-1)$	0.0007	0.0219%
	4	$[u_3(k-1)]^2$	-0.0013	0.0352%
	5	$u_3(k-1)$	0.2844	0.0275%
	6	$[u_2(k-1)]^2$	-0.0004	0.0007%
	7	$y(k-1)u_2(k-1)$	0.0006	0.0026%

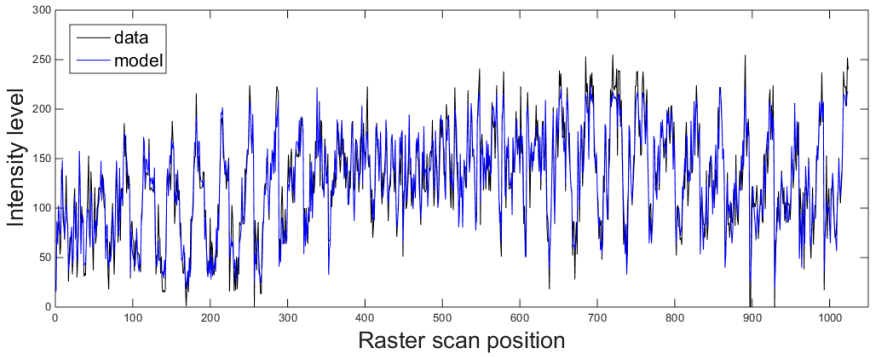


Figure 3.16: Model fitting to data, healthy subimage from mdb009 [212].

Table 3.9: Model of a malign ROI (64 × 64 px, mdb028-182 [212]).

Subimage	Index term	Model term	Parameter	ERR
	1	$u_3(k-1)$	0.6592	99.9102%
	2	$u_2(k-1)$	0.6039	0.0221%
	3	$[u_2(k-1)]^2$	-0.0002	0.0025%
	4	$u_1(k-1)$	-0.2818	0.0006%
	5	$[u_1(k-1)]^2$	0.0007	0.0004%
	6	$[u_3(k-1)]^2$	-0.0003	0.0004%
	7	θ_0	-1.0989	0.0001%

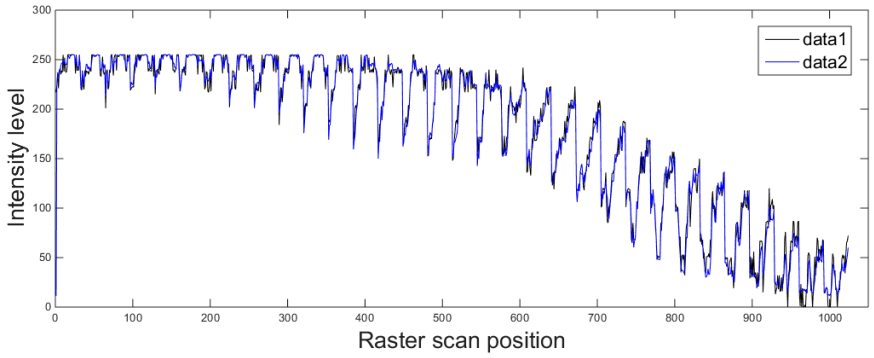


Figure 3.17: Model fitting to data, malign ROI from mdb028 [212].

Since the proposed feature extraction and classification method is designed to analyse at subimage zoom level, it is relevant to present NARX subimages models to illustrate and understand what the models consist of and how do they vary from case to case. In the model fitting to data plots we can observe that the models' capacity to adjust to the learning data was high.

By comparing full mammogram models versus ROIs models, the second group of models contains, mathematically speaking, more distinctive features such as a particular model structure, a higher nonlinearity level and more diversity of candidate model terms in the final solution than the first group. This finding indicates that in breast cancer detection, 64×64 subimage models can better capture more image features because they link together more elements, are much more diverse among themselves and are more responsive to intensity changes than further away zoom image models.

Regarding the analysis of the case studies presented above, we can see that the benign model has very little dependence on non-linear terms, although there is a high presence of exogenous inputs. Its adjustment to real observations is in

practice excellent, where there are data fluctuations from medium to light-grey intensity tones.

The model from a healthy subimage, instead, showed an important model term diversity but a higher dependency on non-linear terms and a consistent presence of autoregressive terms. Its adjustment to observations was high, where data fluctuation from dark to medium grey intensity levels were present. The malign model had a balanced presence of linear and nonlinear terms, but it was entirely dependent on exogenous inputs. Its adjustment to data was also high, where clear to medium grey tones became darker (lower intensity values) at the end of scan position.

Although the data-overfitting is usually a negative factor in predictive models, nearly always designed at having good generalisation properties, the context of feature value extraction allows contemplating the accurate adjustment of models to the data as highly desirable, since it contributes to good image representativeness.

3.3.5 Classification performance metrics

For the sake of feasible comparisons and more interpretable results, the measures of decision performance selected for this study are accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (PNV). A detailed description of these metrics is reported in Section 2.4.2.

3.3.6 Classification and detection results

With the database partitioned into 222 and 100 images for training and testing, as noted in Section 3.3.1, the initial testing results are summarised in Table 3.10. These results only represent the count of hits and mistakes of the 2D-NARX.

Table 3.10: Partition data and initial 2D-NARX results.

Testing database partition	
Instance	**Counting**
Total mammograms	100
Documented positive cases	43
Documented negative cases	57
Classifier hits and mistakes	
Instance	**Counting**
True positives	40
False negatives	3
True negatives	51
False positives	6

The results related to classification performance metrics are summarised later in Table 3.11. Finally, a comparison of the decision performance between the new method and previous parametric models for breast cancer detection is displayed in Table 3.12, where, although the same metrics are used, most reported methods use different databases so a cautious approach is recommended.

The assessment of the method's capacity to discern among benign and malign tumours took place separately. 240 ROIs were chosen for testing through a pondered random sampling and according to the original distribution, with all ROIs containing fully or partially the abnormality, according to the image-coordinates and approximate radius listed in the mini-MIAS documentation. The method accuracy for this test was 94.16%. As for the runtime, the average processing length per subimage was 1.7 seconds.

Table 3.11: 2D-NARX tumour detection results.

Performance measure	**Result**
Accuracy	91%
Sensitivity	93.02%
Specificity	89.47%
PPV	86.96%
NPV	94.44%

Figure 3.18 exemplifies 4 cases made through the 2D-NARX image characterisation to ease a visual appraisal of the new method. Upper subimages are positive or negative cases identified by the new method. Lower ROIs belong to samples of the training set which helped the classifier to *attract* the corresponding image above each. All examples are actual ROIs/subimages of 64 × 64 pixels extracted from mammograms of the mini-MIAS database [212].

Table 3.12: Comparison of 2D-NARX with previous parametric methods.

Model	**Reference**	**Image set**	**Acc. %**	**Sens. %**	**Spec. %**
1D-ARMA	[10]	U. of Illinois	78.5	59.5	79.7
2D-ARMA	[10]	U. of Illinois	93.8	92.3	94.1
2D-ARMA	[11]	DDSM	96.5	96.9	97.8
AR-QUS	[12]	Sunnybrook H.	83	88	91
ELM	[7]	mini-MIAS	91	90	98
2D-NARX	[217]	mini-MIAS	91	93	89.5

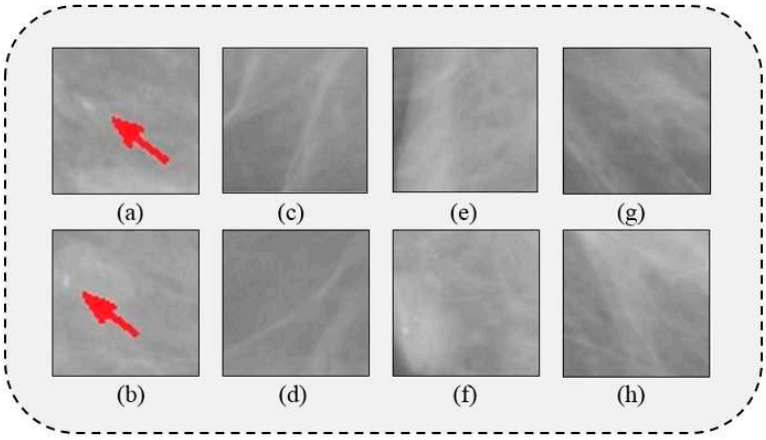


Figure 3.18: Microcalcification (a) falling into the benign class thanks to sample (b). Subimage (c) falling into the healthy class thanks to healthy sample (d). Malign tumours (e) and (g) falling into the malign class tanks to samples (f) and (h) respectively (images from [212]).

3.4 Discussion

- This work presents a new method that identifies linear and nonlinear features from digital images to produce useful information to classification and detection aims. The model design aimed at converting the image data into the input-output system format in the first place. This step included tests with different neighbourhoods and raster scan step sizes. Such drawback came from the approach novelty and the omission of many authors in describing this step in detail.

- For the first time, the NARX model as a system identification method for digital images and as a CAD procedure for the classification of the medical condition and the detection of abnormalities in digital mammograms.

- The use of nonlinear system identification models with flexible structure and model order within image processing enhances the spectrum of CAD systems for breast cancer detection based on parametric models. However, the algorithm is not limited to mammogram analysis or medical image analysis.

- Experiments with the mini-MIAS database of mammograms revealed that the new method produced ad-hoc model structures while included nonlinearities in the image models for the sake of a richer representation.

- As to classification metrics, the algorithm showed a superior sensitivity (93%) but lower specificity (89.5%). These values mean that the failure of noticing tumours was low, but false alarms could be produced together with unnecessary expenses by follow-up examinations. The PPV of 87% indicates that if a positive case is detected, an additional examination is recommended to confirm the suspicion. NPV of 94.5% points out that negative results (normal condition diagnosis) are safer to regard by the medical staff.

- As regards the ROIs, the new algorithm was capable of discriminating benign from malign samples (94%) despite a significant similarity between these two classes in many ROIs. The recommended method did not attain the highest numbers, it stood competitive against previous algorithms getting a high sensitivity, a below-average specificity, and a standard accuracy.

- The adaptive NARX modelling showed to be capable of avoiding identification traps, as the estimation of too simplistic models from images containing many information. Figure 3.5 explains the raster scan strategy to avoid such problem by fully constraining overlaps between consecutive scan positions. For instance, this strategy would prevent the creation of the overly simplified model $y(k) = y(k-1)$ to describe a complex system. That model would

imply that to predict the value located at any position $y(k)$ it would only be necessary to know the information at $y(k-1)$, which is rather a plain model without any descriptive value.

- The supervised learning process involved the human labelling of a large quantity of data, entailing a high cost concerning the person-hours spent and extending the development time estimated initially for this project.

- The fact of carrying out an unsupervised training can easily lead to confusing or misleading classification results, which in medical diagnosis could mean a severe problem that can risk human lives. Given such a scenario and for the sake of increasing the classifier sensitivity, a necessary but time-consuming effort was made during the data labelling to include as possible malignant and benign samples not only from different mammogram zones around ROIs but also from mammogram rotations.

- The recommended algorithm presented limitations and areas of development such as the mistaken, although rare, identification of identical NARX models for two different subimages. To minimise the problem, the image characterization added rotation angles to enrich the feature extraction process. Besides, the new method did not leverage the model structure information, an appealing system identification attribute, to the feature extraction process.

- The above limitations took place in favour of a newly developed identification design that promoted the creation of regular-sized feature vectors. This achievement based on the use of different fixed signals focused on producing model responses to be processed by statistical inference.

- Future work includes the feature vector enhancement through a reselection of statistical estimators, an in-depth exploration of the NARX model set-up

since higher order terms and more input variables can help to enrich the image models, the inclusion of additional preprocessing techniques to normalise uneven contrast levels between mammograms without altering the essential image information and the design of strategies that best map the NARX model structure since it has not been previously seized in image characterisation and can represent a competitive advantage.

Chapter 4

Image classification by MSRBF networks and DCT

The previous chapter presented a new approach to image feature extraction via the NARX model to improve the scope of nonlinear system identification to tackle the breast cancer detection problem. The work presented in this chapter aims at extending the above concept by using the multiscales RBF networks, a nonlinear system identification technique that is powerful but never before used in image processing. The discrete cosine transform is incorporated to characterise the image model and retrieve feature values of reduced dimensionality and high representativeness. Classification results show that the new method reached a very competitive diagnosis accuracy.

The highlights of the proposed method are:

- A 2D NARX image mapping and its adaptation to multiscale radial basis function networks.
- The solution of the MSRBF network using the FROLS algorithm.
- A DCT-based feature extraction process for enhancing the image characterisation.

- The creation of a CAD system based on the MSRBF DCT framework.

Section 4.3 reports the classification test results in a public database of digitised mammogram films.

4.1 Introduction

Digital image processing and computer vision techniques encompass an increasing variety of approaches to real-life problems. When it comes to image classification, image processing methods aim at recognising both visible and hidden patterns to enable a subsequent statistical inference process, oriented in the first place to extract feature values to feed such analytic process [1]. Among the last ones, there is increasing acceptance in the literature on system identification approaches, which are mainly focused on building models only based on the historical record of the system's inputs and outputs [88]. This kind of models is also capable of recognising and reproducing behavioural patterns from a system's behaviour without prior knowledge of its inner structure. Such pattern recognition capability is what makes system identification models highly appealing in image processing. Figure 4.1 shows the core of the system identification scheme.

Computer-aided diagnosis (CAD) is a field of intense development that bridges image processing and computer vision disciplines to the medical field, especially in visualisation and diagnostic tasks. CAD has made the most of the current advances in intelligent systems. Examples are software supporting platforms for radiologists in decision-making [34]. One of the most popular system identification approaches in CAD systems is represented by artificial neural networks (ANN) given their excellent modelling capacity.

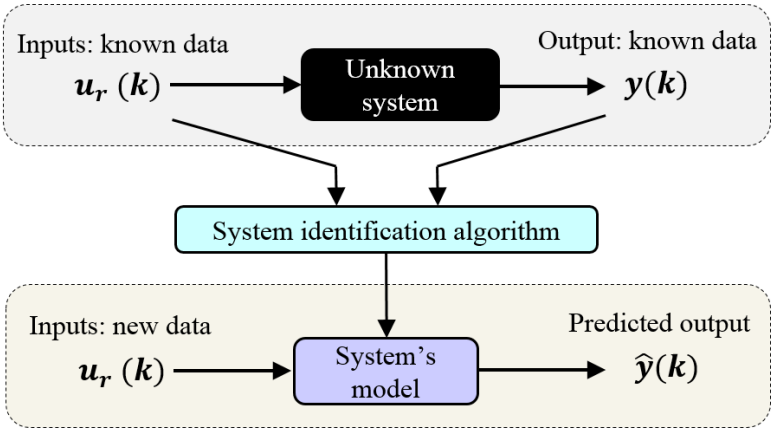


Figure 4.1: Predicting the unknown system's behaviour via system identification.

Moreover, many experts in CAD systems rely more frequently on the use of multi-layered ANN with the intention of obtaining even better approximations. However, the more the hidden layers are included in the network, the slower and more complex the model training becomes. Conversely, single hidden layer networks, as radial basis functions, are known to be sufficient to estimate any continuous function independent of the linearity degree [88],[155].

Radial basis functions (RBF) are popular kernel-based networks which represent a particular class of ANN. Kernels are mathematical functions contributing together to simulate a higher dimensional space from another one of lower dimension to ease the adjustment of relationships between the data by expressing it in a new way. The kernel methods were incorporated into neural networks by 1990s through SVM to solve machine learning-related problems such as classification, regression and object recognition [218]. The popularity of kernels spread later to principal component analysis (PCA), which solves problems with centred data (or zero mean) and is focused on features or principal components of variables linked to the input space [219]. Another popular variation is the polynomial kernel. It

uses the information contained in a group of monomials by initially extracting from these the product features to feed the learning algorithm subsequently [220].

As to probabilistic models, the kernel method incorporated Gaussian processes, also named radial basis function kernels, the most commonly used at present [219]. In spite of some kernels have specific applications, there is not a universal choice for all problems. However, Gaussian kernels have shown to outperform other kernel alternatives in classification problems [221]. Besides, RBF kernels hold a linearly weighted structure that eases the training and discards more complex nonlinear procedures in the solution algorithm, so they are efficient at solving nonlinear system identification problems [155].

Note that nonlinear image analysis has proven to be necessary for an increasing number of areas, and digital image processing is not an exception. Exclusively linear procedures in images may lead to poor operational results regarding edges, non-Gaussian noise and other random distortions, factors that can be especially dangerous when a high accuracy analysis is needed [4],[5],[6].

Notwithstanding that RBF networks sound like a good choice due to their power of modelling and solving simplicity, the approximations they produce may lack the flexibility to model highly dynamic or rapid changing systems. An alternative to this limitation is the multiscale version of RBF, termed Generalized Multiscale RBF networks (MSRBF) that provide a trade-off among the modelling straightforwardness of RBF networks and the advantages provided by more complex deeper networks [222].

Until the presentation of this work, MSRBF networks have not been used in image processing techniques and even less in CAD systems. In this work, the MSRBF networks philosophy is adopted and combined with the discrete cosine

transform to extract high-quality information from images with classification purposes. Moreover, this chapter proposes the NARX mapping of digital images to an autoregressive input-output system format (explained earlier in Chapter 3) with the purpose of making the digital image information consistent with nonlinear system identification problems. Tests results show that the new method is very competitive as a CAD system in breast cancer image detection, an important and challenging public health problem.

Previous work on RBF networks is abundant and a review of the most representative and related techniques was carried out in Chapter 2. In short, RBF networks have been used in general applications as the prediction of near-earth geomagnetic field [15], face recognition [83], [159], modelling and identification of dynamical systems [158], three-dimensional object recognition [160], and motor systems control [161], and in CAD systems involving pathological brain detection [163] and breast cancer detection [7],[162],[8],[9],[204].

This work puts forward a novel image processing framework for feature extraction based on an improved version of RBF networks, adding to it the advantages of the DCT information compression and adapts the new methodology successfully to CAD systems for breast cancer detection. This chapter describes the information flow and the logic behind the proposed method, including all adopted algorithms. Section 3 shows the experiments and results of the methodology. Finally, Section 4 presents a discussion on the findings, difficulties and future work.

4.2 The MSRBF DCT methodology

The MSRBF DCT feature value extraction method bases its logic on four main algorithms: conversion of the image data into the NARX format, the multiscale version of RBF networks, the FROLS algorithm and the discrete cosine transform.

The adjustment of the new methodology into the CAD point of view involved the image partition into subimages or regions of interest (ROIs) in the first place. As in chapter 3, ROIs are regarded here as the standard processing units, where a 64 x 64 pixel-size was assigned to better enclose the ROIs such as tumours and microcalcifications including the surrounding regions. Besides, a splitting process was included to deepen the analysis scope of this work as for the objects' position detection in the ROI area and to produce a two-fold and parallel characterisation (Figure 4.2). In the figure, a complete subimage is observed on the left side, followed by its dual partition on the right. the new two-fold ROI characterisation aims to improve the ability of the framework to retrieve the size and object position from the image more effectively since it helps to allocate information from an image zone into a feature vector section.

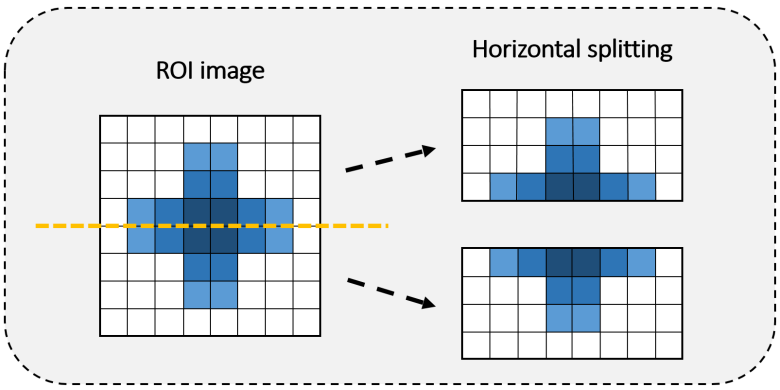


Figure 4.2: ROI splitting for a two-fold characterisation.

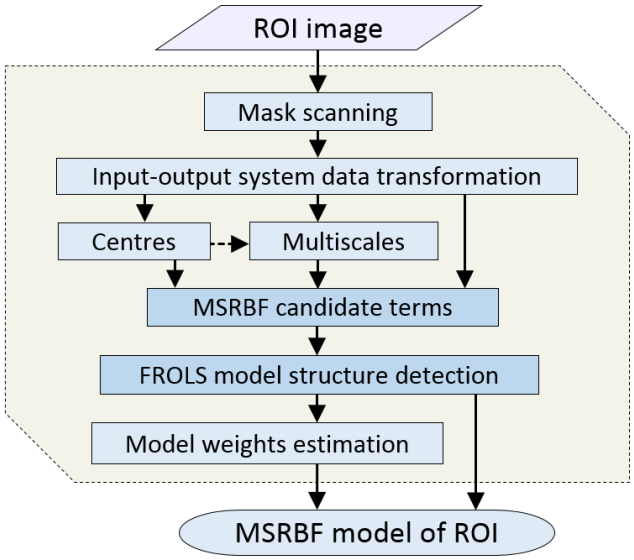


Figure 4.3: MSRBF model approximation flowchart.

As for the image processing, each ROI split is read and stored according to the input-output system format at first. Then, such data must be processed to derive a convenient number of data centres. The referred centres represent artificial neurons or functions contained in the singleton hidden network layer. In this work, the mathematical structure of each neuron-function is the standard Gaussian function, defined alongside the complete processing in the following section.

At the end of the image modelling process, the structure selection algorithm FROLS comes into play to assess the candidate neurons and include the most representative terms into the model. The solution of a system of linear equations of the form $Ax = B$ for x, where A equals the selected terms, x is the vector of parameters θ, and B is the vector of the output y, is processed to obtain the parameters or weights w of the model. Figure 4.3 shows the MSRBF network information flow within the new methodology.

Once the model is available, a set of input signals is used to excite the model and generate a corresponding output signal series, whose values are processed via the DCT and assembled to obtain a feature vector. The same process is repeated with all mammogram's ROIs to compare the final vectors to pre-tagged samples corresponding to healthy, benign or malignant class utilising a distant-based classification algorithm.

4.2.1 Discrete-time system structuring

At this stage, the method aims to scan image data similarly to time series, where instead of discrete time periods, adjacent pixel neighbourhoods lay distributed along the image. From the system identification perspective, the way of representing such data must be congruent with the following equation:

$$y(t) = \hat{f}(\mathbf{x}(t)) + e(t) \tag{4.1}$$

in which the output $y(t)$ is explained by a nonlinear function $\hat{f}$ and an error sequence $e(t)$. Based on the 2D-NARX model describing a single-input-single-output (SISO) system, the nonlinear function is compound together by a list of input-output regressors as follows [217],[88]:

$$\begin{aligned} y(t) =& \hat{f}[y(t-1), y(t-2), \cdots, y(t-n_y), \\ & u(t-d), u(t-d-1), \cdots, u(t-d-n_u)] + e(t) \end{aligned} \tag{4.2}$$

where $\hat{f}$ is an unknown nonlinear function, $y(t)$ is the sequence of the system output, $u(t)$ is the sequence of the system input, n_u and n_y are the maximum lags for the system inputs and output (set up in this work equal to 1), and d is a time

delay auxiliary value, set here to $d = 1$. According to the more complex NARX representation for a multiple-input-single-output (MISO) system seen in Chapter 3, aimed at producing a richer feature extraction, the vector $\mathbf{x}(t) = [x_1(t), \cdots, x_d(t)]^T$ is defined as a set of regressors in the following way:

$$x_b(t) = \begin{cases} y(t-b), & 1 \leq b \leq n_y \\ u(t-(b-n_y)), & n_y + 1 \leq b \leq n_y + n_u \end{cases} \tag{4.3}$$

where n_u and n_y are the maximum lags for the input u and the output y respectively and b is an auxiliary value. Taking into account the previous description, the vector of basic cross-coupled regressors to be combined in the 2D case within the function expansion of this chapter is defined as $x_b(i,j)$. The last representation shift aimed at addressing the bi-dimensional image processing problem of this chapter, compared to the simpler time series problem that depends on a single variable. Please note that Section 3.2.2 details the two-dimensional modelling of this process. With this in mind, Equation 4.3 defines the set of regressors of as follows:

$$x_b(i,j) = \begin{cases} y(i, j-2b), & 1 \leq b \leq 1 \\ u_1(i-1, j-1-2(b-1)), & 2 \leq b \leq 2 \\ u_2(i-1, j-2(b-2)), & 3 \leq b \leq 3 \\ u_3(i, j-1-2(b-3)), & 4 \leq b \leq 4 \end{cases} \tag{4.4}$$

where the maximum lags $n_y, n_{u1}, n_{u2}, n_{u3}$ were fixed in 1 and b is an auxiliary value, following the actual model set up to be seen in Section 4.3.

4.2.2 Traditional RBF and 2D MSRBF neural networks

Traditional RBF neural networks

Traditional RBF networks are known to be straightforwardly structured, but with a considerable power to identify a whole range of systems, including those with irregular data [223]. However, single-scale RBF networks may have modest generalisation qualities [222]. Generalised multiscale radial basis function networks (MSRBF) provide a favourable trade-off between easy to solve traditional RBF networks and the modelling advantages of multi-layer networks, which more than often include various hidden layers and involve nonlinear optimisation steps in the solution process [222]. MSRBF networks are multiscale because on the one hand, the kernel function included is Gaussian, and on the other, such Gaussian function has several widths or scales.

As mentioned, the present work includes the Gaussian kernel, for it allows to easily use centres and widths for an added modelling flexibility, as it enables the structure detection algorithm to choose from more options for a better representation. Figure 4.4 exemplifies how the Gaussian neuron-function processes the input data x according to the parameters μ (mean or kernel centre) and σ (standard deviation, widths or scales) generating a bell-shaped distribution curve in the output.

The Gaussian kernel is known as a multidimensional universal approximator of functions converting a dimensional space into another corresponding one, but with different dimension (usually longer) that helps to linearly separate any type of input data with non-linear dependencies (like most of the real-life problems) to make features or information easier to extract and interpret by machine learning algorithms. The RBF neural network bases its effectiveness on this advantage and

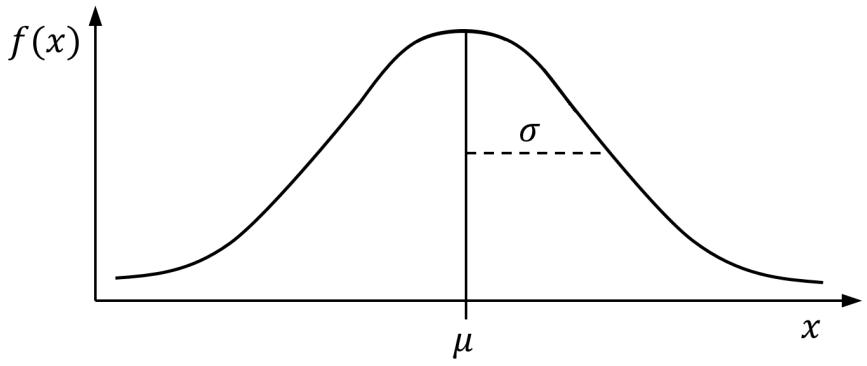


Figure 4.4: The shape of the Gaussian function contained in the RBF kernel.

approximates the unknown nonlinear function $\hat{f}$ utilising a weighted sum of Gaussian radial functions. Figure 4.5 shows the typical architecture of RBF networks.

The RBF structure consists of three layers, where the first one represents the input data linked to the independent variables $x_1, \cdots, x_m$. The first layer is fully connected to the second intermediate layer, formed by the Gaussian neurons

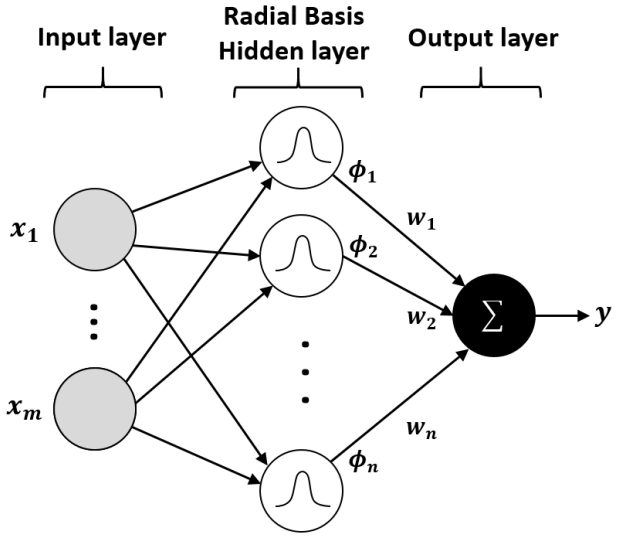


Figure 4.5: Multiple-input single-output architecture of a Gaussian RBFNN before the multiscale expansion to be shown in Figure 4.6

$\phi_1, \cdots, \phi_n$. The second intermediate layer is in turn fully connected to the third layer or output layer, employing the kernel weights $w_1, \cdots, w_n$, which are part of the result of the network training. Note that in the context of the neural network, the Gaussian functions parameters c_i for the centres and μ_i, for the widths are not given in the problem and thus must be computed automatically from data. For this reason, RBF networks are nonparametric methods. The general formulation of the standard RBF for a one-dimensional system is the following:

$$\hat{f}(\mathbf{x}(t)) = \sum_{i=1}^{M} \theta_i \phi_i(\mathbf{x}(t); \sigma_i, \mathbf{c}_i) \tag{4.5}$$

where ϕ_i is the Gaussian kenrnel, $\mathbf{x}(t)$ is the vector of independent variables (which in the NARX model are rather regressors), $\sigma_i = [\sigma_1, \cdots, \sigma_n]$ is the vector of parameters of the scales or widths and $\mathbf{c}_i = [c_1, \cdots, c_n]$ is the vector of parameters of the kernel centres. In such a way, the Gaussian kernel function for a one-dimensional system is stated as follows:

$$\phi_i(\mathbf{x}(t) : \sigma_i, \mathbf{c}_i) = \exp\left[\sum_{b=1}^{d} \left(\frac{x_b(t) - c_{i,b}}{\sigma_i}\right)^2\right] \tag{4.6}$$

where $d = n_u + n_y$, being n_u, n_y the maximum lags for the system input and output and b an auxiliary value for indexing the regressive variables contained in $\mathbf{x}(t)$.

Two-dimensional MSRBF neural networks

The MSRBF network implemented in this new framework adopts the multiscale approach as a primal contribution together with the two-dimensional perspective to tackle the image processing problem. The multiscale extension to RBF, as the name suggests, multiplies the scales or widths of each kernel function with the aim

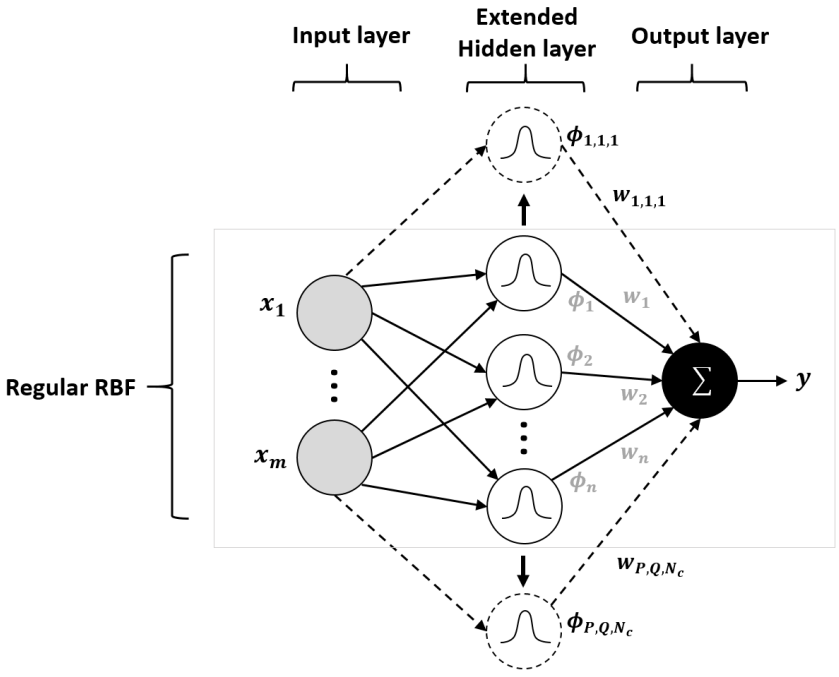


Figure 4.6: Increase in the number of RBF neurons produced by the multiscale approach regarding the architecture shown earlier in Figure 4.5

of expanding the flexibility of the single hidden-layer neural network and better approach the non-linear function $\hat{f}$.

Figure 4.6 describes the structure of MSRBF neural networks, where the vectors of the input layer are fully connnected to the Gaussian kernel functions $\phi_{p,q,m}$ (defined originally in traditional RBF networks as ϕ_i). The number of functions represents the number of kernel centres c_m. The hidden layer neurons are fully connected to the output layer by means of a series of weights $w_{p,q,m}$ corresponding to the model parameters $\theta_{p,q,m}$ stated below in Equation 4.7. The 2D MSRBF version replaces the vector of regressors $\mathbf{x}(t)$ (Equation 4.3) by the two-dimensional vector $\mathbf{x}(i,j)$ (Equation 4.4).

The 2D MSRBF network implemented in this work presents the following mathematical structure, which is an adaptation of a definition presented in [222].

$$y(i,j) = \hat{f}(\mathbf{x}(i,j)) = \sum_{p=0}^{P}\sum_{q=0}^{Q}\sum_{m=1}^{N_c}\theta_{p,q,m}\phi_{p,q,m}(\mathbf{x}(i,j);\sigma_m^{(p,q)};\mathbf{c}_m) \tag{4.7}$$

where $y(i,j)$ is the system output, $\mathbf{x}(i,j)$ is the vector of bidimensional regressors composed of lagged inputs and outputs, $\sigma_m^{(p,q)}$ are the scales, $\mathbf{c}_m$ are the candidate centres with N_c representing their quantity in the network, $\phi_{p,q,m}$ are the basis functions and $\theta_{p,q,m}$ are the model weights to be estimated during training. In that way, the basis functions previously defined in traditional RBFs (4.6), are defined in the 2D MSRBF network as:

$$\phi_{p,q,m}(\mathbf{x}(i,j):\sigma_m^{(p,q)},\mathbf{c}_m) = \exp\Bigg[-\sum_{b=1}^{d}\Big(\frac{x_b(i,j)-c_{m,b}}{\sigma_{m,b}^{(p,q)}}\Big)^2\Bigg] \tag{4.8}$$

where in the same fashion, $\phi_{p,q,m}$ is the general Gaussian kernel, $\sigma_m^{(p,q)}$ are the Gaussian multiscales, $\mathbf{c}_m$ are the Gaussian centres, b is an auxiliary value indexing the variables contained in vector $\mathbf{x}(i,j)$ and $d = n_y + n_{u_1} + n_{u_2} + n_{u_3}$, being n_{u_1},n_{u_2},n_{u_3} and n_y the regressive variables of the multiple-input-single-output network design proposed in this chapter. However, special attention must be paid in the determination of the Gaussian parameters.

Estimation of the kernel centres

The proposed method includes the implementation of an adaptive algorithm to determine the number of centres N_c (and therefore, the number of Gaussian functions in the hidden layer), taken from the work of [222] and [224]. In the first place, *the*

sum-of-squares clustering algorithm acts as a criterion for estimating the number of centres. The algorithm includes the following steps:

1. The input data, composed of N rows and p columns, is divided into an arbitrary number of k initial groups $G_1, \cdots, G_k$.

2. The geometry centre (centroid) $\mathbf{c}_j$ of each group G_j is obtained.

3. The variability d_j per group is estimated by summing all distances of $\mathbf{z}_i$ with respect to the centroid $\mathbf{c}_j$:

$$d_j = 2 \sum_{i \in I_j} \|\mathbf{z}_i - \mathbf{c}_j\|^2 \tag{4.9}$$

   where the vector $\mathbf{z}_i$ is the ith row of input data belonging to the group G_j.

4. The variability function of k, W_k, is estimated by summing the d_j of all groups.

5. The process is repeated from step 1 to 4 using different k values to estimate their variability function W_k.

6. The difference in the variability function of k values involves the following formula:

$$DIFF(k) = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k \tag{4.10}$$

7. The following equation helps to compute the effectiveness of each k by comparing the values obtained in step 6:

$$E(k) = |DIFF(k)/DIFF(k+1)| \tag{4.11}$$

8. Finally, the recommended k value, or number of centres N_c, is that one maximising the function $E(k)$.

After the estimation of the number of kernel centres, the K-means++ algorithm [215] is used to compute a corresponding number of centroids from the $N \times p$ size input data matrix.

Estimation of multiple scales

As for the scales, a two stages determination was carried out, according to the strategy recommended in [222]. The idea behind this aim is to estimate a single scale by basis function ϕ_i in the first place followed in turn by the computation of the quantiles (points taken at regular intervals) resulting from the first scale. Thus, the equations below define the first single scales.

$$\sigma_y = max\{y(i,j)\} - min\{y(i,j)\} \tag{4.12}$$

$$\sigma_{u_r} = max\{u_r(i,j)\} - min\{u_r(i,j)\} \tag{4.13}$$

where σ_y is the initial scale for the output and σ_{u_r} are the initial scales for the inputs $u_r = [u_1, \cdots, u_R]$ of the MISO system. For the calculation of the multiple final scales, the following formula, taken from [222], is used to expand σ_y and σ_{u_r}:

$$\Lambda_m^{(p,q)} = diag\big[\underbrace{(\sigma_{y,m}^{(p)})^2, \cdots, (\sigma_{y,m}^{(p)})^2}_{\text{output y}}, \underbrace{(\sigma_{u_r,m}^{(q)})^2, \cdots, (\sigma_{u_r,m}^{(q)})^2}_{\text{input } u_r}\big] \tag{4.14}$$

where $\Lambda_m^{(p,q)}$ are the covariance matrices for the values $p = 0, \cdots, P$ and $q = 0, \cdots, Q$, u_r are the system inputs and $\sigma_{y,m}^{(p)} = 2^{-p}\sigma_y$ and $\sigma_{u_r,m}^{(q)} = 2^{-q}\sigma_{u_r}$ are the quantiles linked to the output and input initial scales. In this work, the values of P and Q were fixed in 1 and the number of system inputs, R, in 3. Therefore the scales contained in (4.14) can be disaggregated as follows:

$$\sigma_{y,m}^{(p)} = \left[(\sigma_y 2^0)^2, (\sigma_y 2^{-1})^2\right] = \left[(\sigma_y)^2, (\sigma_y/2)^2\right] \quad \text{for all } m \tag{4.15}$$

$$\sigma_{u_r,m}^{(q)} = \left[(\sigma_{u_r} 2^0)^2, (\sigma_{u_r} 2^{-1})^2\right] = \left[(\sigma_{u_r})^2, (\sigma_{u_r}/2)^2\right] \quad \text{for all } m \tag{4.16}$$

where $u_r = [u_1, u_2, u_3]$ corresponds to the system inputs, σ_y and σ_{u_r} are the initial scales for the output and the inputs obtained in Equations 4.12 and 4.13, and m indicates the kernel centre, where $m = 1, \cdots, N_c$. Note that in this chapter N_c equals the number of centres k recommended by the sum-of-squares clustering algorithm, examined earlier. With the above definitions, a more explicit representation of the multiscale radial basis functions expressed in Equation 4.8 is:

$$\begin{aligned}\phi_{p,q,m}(\mathbf{x}(i,j) : \sigma_m^{(p,q)}, \mathbf{c}_m) = \exp\Bigg[&- \sum_{b=1}^{n_y} \Big(\frac{x_b(i,j) - c_{m,b}}{\sigma_{y,m}^{(p)}}\Big)^2 \\ &- \sum_{b=n_y+1}^{d} \sum_{r=1}^{R} \Big(\frac{x_b(i,j) - c_{m,b}}{\sigma_{u_r,m}^{(q)}}\Big)^2 \Bigg]\end{aligned} \tag{4.17}$$

where $d = n_y + n_{u_1} + n_{u_2} + n_{u_3}$ and $R = 3$ is the number of system inputs. After the definition of the kernels, a matrix of candidate functions must be constructed to allow the FROLS algorithm to select the model structure.

4.2.3 Model structure detection

By taking into account that the number of scales in a MISO system is $N_s = (P+1)(Q+1)^R$ and that the model set up in this work was $P = Q = 1$ and $R = 3$, the initial number of Gaussian centres k of this work was scaled $N_s = (2)(2)^3 = 16$ times in the MSRBF network, similarly to the structure expansion shown earlier in Figure 4.6. Following up on this idea, the number of M candidates of the MSRBF network in a MISO system is $M = N_c N_s$, where N_c is the number of kernel centres.

The listing of the M candidates gains importance in the structure detection algorithm since it makes use of a dictionary containing M candidate functions, from which the selection process is carried out. The following dictionary with triple index D_3 enlists the basis functions in the following manner:

$$D_3 = \left\{ \phi_{p,q,m}(\cdot, \sigma_m^{(p,q)}, \mathbf{c}_m) \quad \text{for all} \quad p, q, m \right\} \tag{4.18}$$

where $p = 0, \cdots, P$, $q = 0, \cdots, Q$ and $m = 1, \cdots, N_c$. The forward orthogonal least squares regression (FROLS) algorithm [225] is designed to build, term by term, the best and most concise models by taking into account D_3, the pool of candidate terms. FROLS is initially based on the original OLS estimator [185], which iteratively looks for the candidate terms that best minimise the error respecting the model output $y(t)$ by using the ERR estimator. The orthogonalisation algorithm helps to exclude from the selection the candidate terms which content is redundant to that already included in the model.

However, the ERR estimator in the OLS is biased towards the inclusion of terms sorted first in the model equation [88]. The FROLS algorithm contributes to removing that shortcoming by adding a reordering of the candidate terms within the equation, leaving out biases of any kind in the inclusion of the most significant candidates. Section 2.3.1 gives a detailed explanation of the FROLS algorithm. In this chapter, the stop-criterion of the FROLS algorithm changed to an *IF* function to limit the number of terms. Thus, the model detection ends up when the error tolerance is satisfied or when the model is long enough.

4.2.4 Feature extraction and the DCT

Feature extraction

The feature extraction module of this framework works out from the image models estimated by the MSRBF network. Similarly to the grey box stimulation-response process seen in Chapter 3, in this work a finite number of fixed signals are used to obtain responses from the MSRBF model. However, unlike the 2D NARX model, the featuring process of the model's response signal includes the discrete cosine transform (DCT) to improve the representativeness of the image values concerning the quality and the quantity. This improvement is because the featuring of the model's output response signal takes place through a direct data transformation instead of external measures based on statistical measures, which can be useful but can ignore information when measuring from the outside. Section 2.1.6 and the following subsection explain the basics of the DCT algorithm.

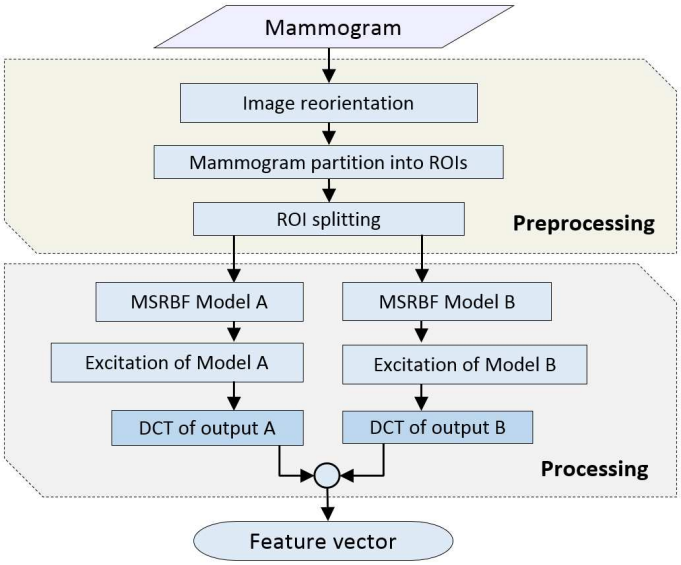


Figure 4.7: Flowchart of MSRBF-based image processing for feature extraction.

Figure 4.7 shows a scheme of the MSRBF-based image processing method, where the excitation of image models and the DCT play essential roles. The idea behind using the DCT is to obtain feature vectors of the same size. Another idea behind is its capacity to allow the choice of an identical number of coefficients per image and therefore create normal feature vectors. Below is a description of this data transformation.

The discrete cosine transform

The discrete cosine transform [80] is a function that computes a sequence of discrete values out of a first sequence. The resulting coefficients are calculated by summing cosine functions valued at various frequencies, producing an oscillating effect in the resulting numbers. A relevant contribution of the DCT is the data compression capability for audio and image processing applications, including pattern recognition [1]. A simple way to explain the DCT is to imagine a vector of a certain length and the DCT as a transformation matrix so that the product of the first two results in a second vector of the same length but with the information concentrated in fewer coefficients. Because of this quality, it is easy to reorder and leave out the less important values. More formally, the DCT for a data sequence $X(i), i = 0, 1, \cdots, (N-1)$ is:

$$F_x(u) = \begin{cases} \frac{\sqrt{2}}{N} \sum_{i=0}^{N-1} X(i), & u = 0 \\ \frac{2}{N} \sum_{i=0}^{N-1} X(i) \cos \frac{(2i+1)u\pi}{2N}, & u = 1, 2, \cdots, (N-1) \end{cases} \tag{4.19}$$

where $F_x(u)$ is the ith DCT coefficient. Figure 4.8 illustrates a graphic example of a 2D DCT compression . In the example, it is possible to appreciate the information compression effect to only a few values of the image, compared with the original

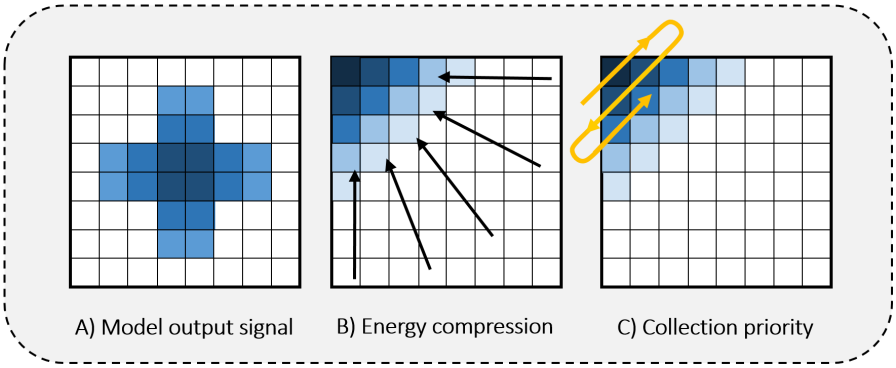


Figure 4.8: Example of 2D DCT information compression in a ROI.

image. This same effect applies to the analysis in one dimension, where the first few coefficients concentrate the resulting information compression.

4.2.5 Classification and detection

The classification module is the connection between the feature extraction process and CAD systems. It links the feature vectors from the supervised, pre-labelling

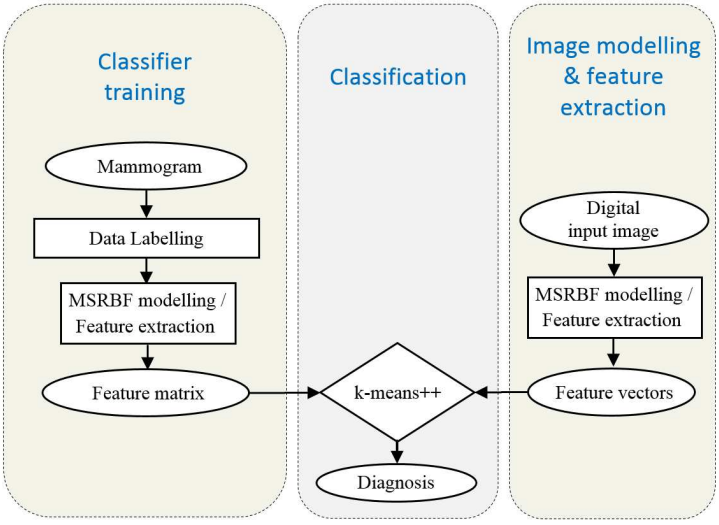


Figure 4.9: Role of the MSRBF DCT into a classification-based CAD system.

task with the unlabelled feature vectors of the image to classify according to the case study's classes. Figure 4.9 aims to ease the information flow visualisation of the proposed framework. In the chart, we observe two separate parallel processes of image data extraction converging into the detection/diagnosis module, based on classification. The difference between classification and diagnosis is that the first one associates the input vector with a class. The diagnosis module uses the classification results to interpret the patient's condition and displays a message easy to understand. For classification, the distance-based K-means++ algorithm was selected [215]. The standard algorithm K-means inspired this technique. However, K-means++ holds the advantage of using an improved seeding method to choose centres, producing an efficient classification up to 70% faster [215].

4.3 Experiments and results

The assessment of the MSRBF DCT method engaged various experimental steps. The chosen repository was the mini-MIAS database of mammograms [212]. The public repository includes 322 high-quality grayscale X-ray films of 1024×1024 pixels of the medio-lateral oblique view of the breast in PGM format. The evaluation goal was to assess the quality of the feature extraction method by evaluating its classification quality for a defined set of mammograms with information attached to them regarding the medical condition class and the background tissue type. The database distribution regarding the breast tissue type is detailed in Table 4.1.

A randomised data-splitting of the 322 breast scans of the database was made, following a 65% to 35% ratio for training and testing with the aim of reducing the chance of attaining biased performance metrics. Furthermore, to counteract the

Table 4.1: Database breast-type distribution [212].

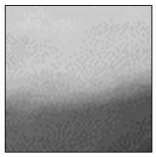
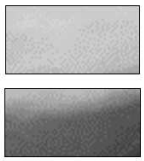
	Fatty	**Fatty-Glandular**	**Dense**	**Total**
Count	106	104	112	322
Percentage	32.92	32.3	34.78	100

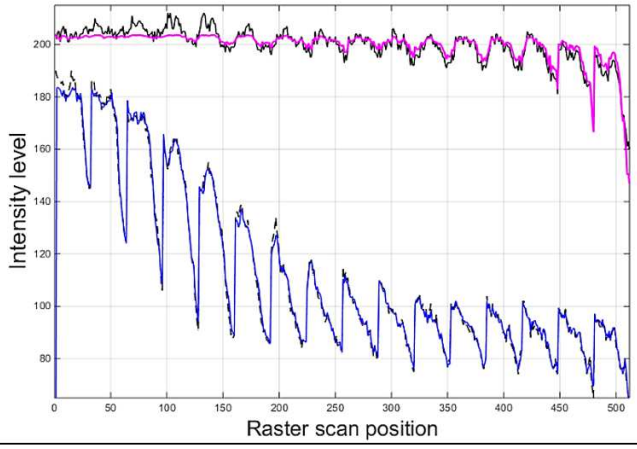
high image variability regarding the breast tissue type, $n = 4$ different training and testing scenarios with different tissue background composition were carried out aimed at, on the one hand leaving in evidence potential differences in the classification results and on the other to get a set of final performance measures with minimal bias.

In that way, the global accuracy for a n number of training and testing scenarios is defined by $Accuracy_n = average(Accuracy(i))$, where $i = [1, \cdots, n]$ symbolises the ith test. All programs were coded in MATLAB R2014b 64-bit and executed in a computer running the Windows 7 Professional operating system with Intel (R) Core (TM) i5-4590 processor at 3.30GHz speed, running MATLAB 2014b.

The assembly of a matrix of 21,637 feature vectors for data labelling produced 95.5% of vectors belonging to the normal class, and 4.5% identified as abnormal, being 2.29% benign and 2.21% malign. The error tolerance of the ERR stop-criterion was 0.15%, and the maximum number of terms was 2. Once the mammogram partition into ROIs (divide and conquer) was done, the feature vector extraction and feature vector labelling of the complete database took place by processing the subimages into vectors and matching the database documentation with these vectors. After the full database characterisation it was possible to build any feature matrix for a specific training partition through the creation a subset

Table 4.2: Two pairs of fit-to-data curves and ERR values. ROI from [212].

ROI	Image split	ERR	Legend
		99.78%	Data Model
		99.82%	Data Model



of the entire matrix database via the removal of the mammogram-related vectors selected for testing.

The initial evaluation aimed at judging the ability of the model to fit the observational data. Table 4.2 shows the example of a dense tissue-type subimage or ROI, its subdivisions (for a two-fold characterisation) and the error reduction ratio (ERR) of the models concerning the data of each case. The table above includes a plot overlying the fit of both models versus the original data. It is possible to observe from the chart that the model adjustment is reliable in both cases since the curves of the predicted output and the original data overlap each other in both pairs of curves.

To expand the evaluation of the MSRBF DCT image model to forecast the observed data, five performance measures conventionally used in machine learning

Table 4.3: Indices for precision validation for the two pairs of model prediction-to-data curves displayed earlier in Table 4.2.

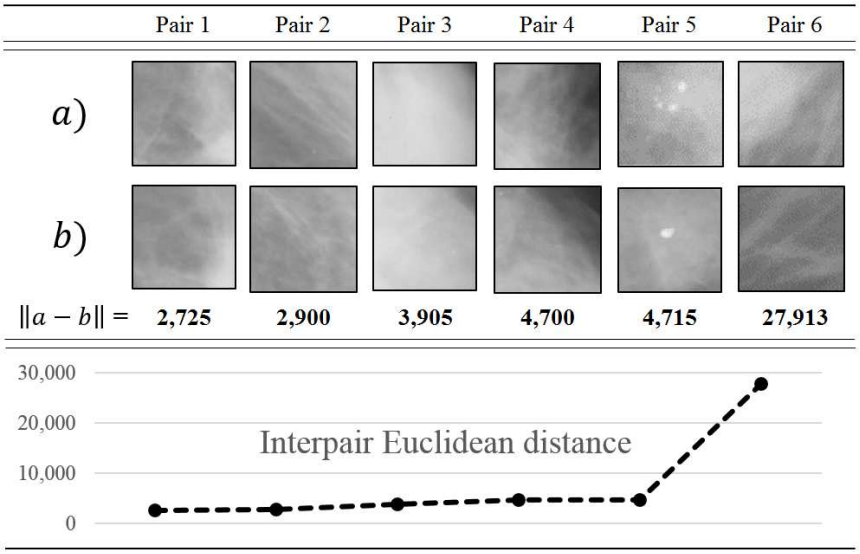
	MSE	**MAE**	**MAPE**	**RMSE**	**NRMSE**
Upper ROI	7.2246	2.097	1.06	2.6879	0.0517
Lower ROI	66.4114	1.8062	1.4843	8.1493	0.0613

and forecasting were estimated and presented in Table 4.3, which incoporates the Mean squared error (MSE), the Mean Absolute Error (MAE), the Mean Absolute Percentage Error (MAPE), the Root-Mean-Square Error (RMSE) and the Normalised Root-Mean-Square Error (NRMSE). The formulations and in-depth analysis of the above measures can be found in [226] and [227].

The MSE measures the prediction quality using the square average of the deviations, where errors are highly penalised. The RMSE is the most used metric in regression models and uses the standard deviation of the prediction error. It amplifies significant errors and is more robust than MSE. The MAE averages the absolute values of the prediction errors and is easier to interpret than the RMSE. The MAPE provides an intuitive interpretation by using the percentage of error between prediction and data. The NRMSE is the standardised version of RMSE and is ideal for comparisons because it is more robust to unit changes.

From the tables 4.2 and 4.3 some observations can be made. In general, the deviation values in both settings are low, taking as reference the original scale where the intensity value goes up to 210. Visually, the adjustment of the two curves is quite good, but the values of MSE and RMSE penalised the Lower ROI prediction strongly, with values of 66.4 and 8.1 compared with 7.22 and 2.68 for the upper ROI. Conversely, the MAE value indicates that the deviation of the upper ROI is

Table 4.4: Six ROI pairs, each aligned vertically, and below them the Euclidean distance between their feature vectors. These vectors are obtained from the image model output's DCT compression. Note that the more the visual difference, the larger the gap. Images from [212].



the highest, which suggests that the lower ROI adjustment is rather trustworthy. With the above, it can be inferred that the MSRBF DCT model makes reliable predictions in general, although its accuracy can decrease when sudden changes in intensity level occur, such as those of the low ROI image. After this point, the feature extraction assessment focused on the consistency of the Euclidean distance between pairs of feature vectors. The Euclidean distance is the length in the space between two points, say $\mathbf{a}(x_1, y_1)$ and $\mathbf{b}(x_2, y_2)$ in a straight line. It can be defined for points $\mathbf{a}$ and $\mathbf{b}$ as follows.

$$d_2(\mathbf{a}, \mathbf{b}) = \sqrt[2]{|x_1 - x_2|^2 + |y_1 - y_2|^2} \tag{4.20}$$

The Euclidean distance relationship was expected to be proportional to the visual image similarity between ROIs. Table 4.4 exemplifies the comparison exercise. The chart represents six pairs of images (a) and (b) holding different separation degrees. Below the images are the interpair Euclidean distances between the image vectors. At the bottom, there is a curve showing how the gap increase as the image pairs display a more significant disparity.

The experimental performance results of the four tests from different database partitions are described in Table 4.5. The classification metrics detailed in Section 2.4.2 supported the assessment of the new model. At first, the percentages by mammogram-type included in each test are displayed. The overall results are quite encouraging in the four tests, especially regarding accuracy, specificity and NPV. As assumed, we can note that the test set composition impacted the classification

Table 4.5: MSRBF DCT performance results by breast tissue-type ratio.

		Test 1	**Test 2**	**Test 3**	**Test 4**
Tissue ratio	Fattty %	31.86	31.86	38.05	34.51
	Dense %	29.20	31.86	28.32	23.89
	Glandular %	38.94	36.28	33.63	41.59
Performance	Accuracy %	93.81	91.96	93.81	94.69
	Sensitivity %	85.00	87.50	87.80	87.88
	Specificity %	98.63	94.52	97.22	97.50
	PPV %	97.14	89.74	94.74	93.55
	NPV %	92.31	93.24	93.33	95.12
	Lesion distinctinon %	81.97	80.88	74.55	76.00

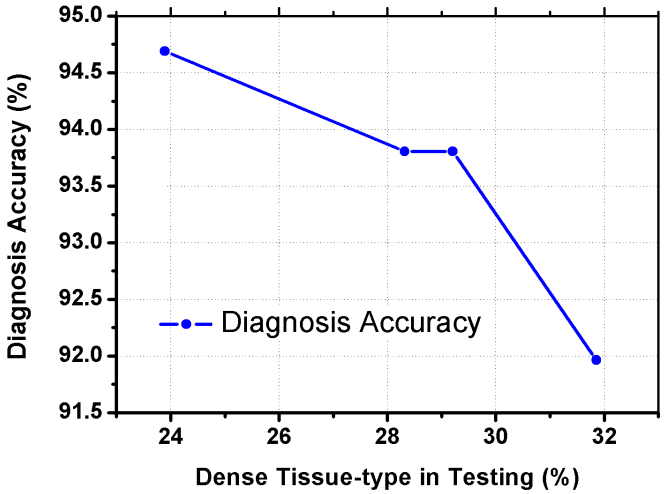


Figure 4.10: Accuracy as a function of the presence of dense mammograms in the test set.

results. This finding is an interesting point to discuss as this factor may lead to confusing results in the breast cancer detection problem.

Note here that the sensitivity and specificity reported in Table 4.4 contrast with those in Table 3.11 of the previous chapter, even though the same database was used. As mentioned later in the chapter closure, this is because, during the MSRBF DCT model training, four times more samples were taken into account, so the sampling strategy was different. Also, the training approach aimed at reducing a constant tendency of the model to find false positives, due to a common difficulty of the breast cancer detection problem related to healthy but quite dense samples.

To ease the analysis of the resulting variations of the classification concerning the mammogram-type composition in the testing set, exciting trends in the results were found and plotted.

Figure 4.10 shows a negative relationship found between the presence of dense mammograms in the test set and the classification accuracy. Such divergence can

Table 4.6: False (positive and negative) cases during testing. ROIs from [212].

	Case A → F/P	Case B → F/P	Case C → F/N	Case D → F/N
Labelled ROI				
ROI bound by the classifier				

be the result of that dense-healthy images are visually similar to tumours of high density, producing false detections.

As an example, Table 4.6 shows four cases: two false positives (false detections) and two false negatives (erroneous omissions) produced during testing when ROIs with dense or glandular tissue were involved. It is possible to see that the abnormal tissue and the dense-healthy or glandular-healthy tissue can come to have quite similar image compositions, causing, therefore, potential erroneous links by the classifier.

On the other hand, Figure 4.11 suggests that there was a lessening ability to distinguish the abnormality class with the increase of fatty mammograms presence in the test, which was opposite to the expected result, given that fatty tissue tends to have translucence, which would make the classification procedures easier.

However, and in favour of the latter hypothesis, the change of the sensitivity values in the different set compositions suggested a positive trend between fatty tests sets and the effective detection of any abnormalities (benign or malign). The last point, together with the accuracy decrease in denser compositions, led to finding a positive relationship between the MSRFB DCT classification accuracy

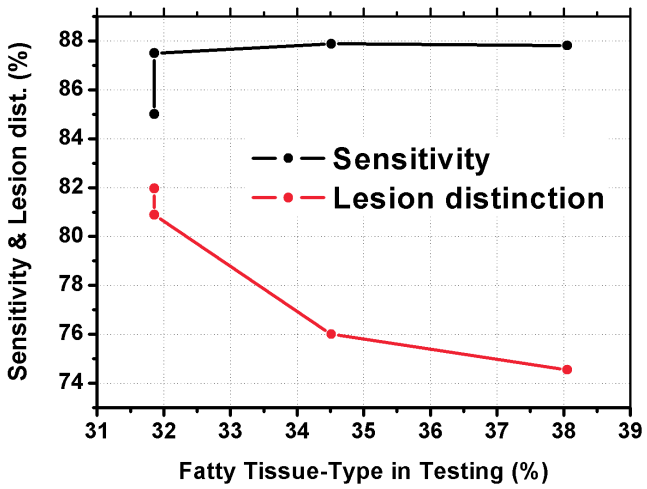


Figure 4.11: Sensitivity and lesion distinction accuracy as functions of the presence of fatty mammograms in the test.

with fatty mammograms and a negative relationship with dense mammograms. Although these results may seem intuitive, it is necessary to carry out more discriminative studies of this type in the future, especially with other methods of featuring and classification to draw more generalised conclusions regarding the breast cancer detection.

As for the variation of the presence of glandular mammograms in the testing set, Figure 4.12 shows a very light direct relation of specificity and NPV with the presence of glandular tissue. Although the trend was not significant enough to be taken into account, it was expected, however, that the presence of glandular tissue, on the contrary, would actively impede the quality of the classification results.

The overall performance of this study is presented in Table 4.7. It is noticeable that values of sensitivity, PPV and lesion distinction are not as high as expected, possibly because of the high resemblance of dense-healthy and glandular-healthy tissue with many abnormal tumours. Among all the values, it stands out

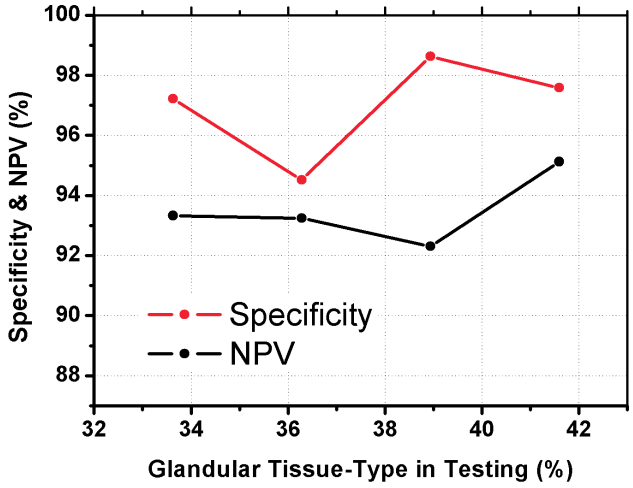


Figure 4.12: Specificity and NPV as functions of the presence of glandular mammograms in testing.

that the tumour distinction was the lowest value of all the registered ones. This difference is due on the one hand to the relative scarcity of abnormal samples, which represented only 4.5% of the total of the labelled samples and on the other to the fact that machine learning methods are generally more efficient to a more significant number of samples available for training [228].

Table 4.7: MSRBF DCT overall performance results.

Statistical Measure	**Average result (%)**
Accuracy	93.57
Sensitivity	87.05
Specificity	96.97
PPV	93.79
NPV	93.50
Lesion Distinctinon	78.35

Table 4.8: Comparison of the MSRBF DCT method with previous work.

Model	Reference	Image Set	Acc. %	Sens. %	Spec. %
2D-NARX	[217]	mini-MIAS	91	93	89.5
ELM	[7]	mini-MIAS	91	90	98
GLCM	[162]	mini-MIAS	93.9	97.2	91.5
ICA-RBF	[8]	mini-MIAS	88.2	–	–
LDA-ANN	[9]	mini-MIAS	93.1	99	83
GPZM	[204]	mini-MIAS	89.3	83.5	93.4
MSRBF DCT	[229]	mini-MIAS	93.5	87	96.9

Finally, a comparison of the new method with previous work is presented in Table 4.8. The comparative table shows that the proposed method obtained high accuracy and the highest specificity (the capacity to correctly detect negative cases). However, it showed an acceptable but below the average sensitivity, which placed it below the GLCM method [162], which had the most stable and positive performance.

As observed earlier in the chapter, the writer considers that the limitation mentioned above is the side effect of a contingency strategy during training to reduce the model tendency to find false positives caused by the considerable resemblance between healthy dense tissue and some tumour types. Hence, it became necessary to increase both the number of dense-type healthy samples and benign and malignant tumour samples in the training database, which could have caused an incidental imbalance.

4.4 Discussion

- This chapter presents an advantageous modelling neural network framework originally designed to model nonlinear observational input-output series as a novel image feature extraction method and CAD system, where the DCT algorithm was incorporated to make the most of the MSRBF network image modelling.

- The experiments aimed at appraising the tumour detection in X-ray mammograms showed that the method was competitive compared to well-known previous CAD systems for breast cancer based on system identification and artificial neural networks.

- The proposed method reached a classification accuracy above 93%. While the MSRBF DCT method is not perfect, the below-average classification metrics may be due to a possible faulty data labelling strategy aimed at reducing the high incidence of false positives, added to a frequent similarity found between solid tumours and the healthy-dense tissue.

- A change from the regular FROLS stop-criterion to an *IF* function helped to shorten the computational burden and the runtime in the service of the massive processing of ROIs without a noticeable trace of reduction of the modelling quality.

- The null incidence of identical models for similar images helped to know that the new two-fold ROI characterisation extended the ability of the model to extract size and object position features from the image effectively, which otherwise might be lost.

- As regards comparisons with previous work, care is advisable when interpreting the values, since non-public databases are used in most cases, making it difficult to make a realistic judgement. Besides, no reference about the tissue-type composition included in the test set was available, a factor that is considered capable of producing changes in the global performance.

- The comparison exercise of the model performance with different training-testing compositions allowed to infer that getting results with a single partition in a heterogeneous database may generate unwanted trends in dependence on the percentage of challenging elements such as dense tissue samples.

- Future work includes the transfer of the methodology to other medical study areas such as brain diseases and lung cancer detection. Also, the use of the Receiver Operating Characteristic (ROC) curve could balance the training matrix composition to get to an optimal balance between sensitivity and specificity. The integration of the ROC analysis to the new method could lie in the tailoring of a confidence threshold able to separate positive and negative decisions, such as the modification of a visual criterion during training, so the human can decide the class to which the sample belongs. However, this path could be quite expensive in terms of time with thousands of subimage samples.

- Note that the ROC curve was not considered in this work, as it is plotted from the gradual modification of a decision threshold (namely confidence threshold) between positive and negative cases [230]. The MSRBF DCT scheme cannot easily consider such changes since the decision threshold is not a variable or a value to be modified, but rather it is an undefined, implicit function of the multiple samples collected and tagged in training according to the database documentation.

Chapter 5

Image classification by Multilayer-Fuzzy ELM

This chapter details a new method of digital image classification based on Multilayer Fuzzy Extreme Learning Machine (ML-FELM) networks, where the layers belonging to ELM contribute to the extraction of high-quality feature values, while the last ML-F layers play the role of a classifier with the added advantage of fuzzy logic systems. Given the need to use a ready-to-use test instance during the development process, an image database of handwritten digits is used at first. Later on, the central tests for the problem of breast cancer detection show that the model achieves high speed and classification performance while keeping model simplicity. The method development includes the following points:

- The basics of extreme learning machines and their extension to MIMO RBF neural networks.
- MIMO IT2-RBF neural networks for classification with uncertainties.
- Kernel-based ELM autoencoders for feature extraction.
- The new ML-FELM framework for image feature extraction and classification.

Experiments on the performance of the ML-FELM classifier and comparisons with the previous work are reported in Section 5.4.

5.1 Introduction

The American Cancer Society (ACS) [231] reported that during the last decade breast cancer has been the leading cause of premature mortality and the second cause of death from cancer among women [232]. In 2015, the ACS issued a large number of recommendations for women of different ages to have regular mammography exams as an early detection strategy [231]. Until now, mammography has been an effective visual mechanism for detecting the presence of suspicious masses as benign or malignant [205],[26]. Nevertheless, in mammograms, the low contrast among healthy tissues and lesions makes it hard to distinguish healthy masses from malignant ones. This way, a significant number of efforts to construct intelligent computer-aided diagnosis systems (CADs) have been proposed [26],[233],[205],[229],[234],[235]. In particular, multilayer and deep neural structures have demonstrated to be a promising machine learning tool for medical image processing [236],[154].

In this sense, Multilayer Extreme Learning Machines (ML-ELM) are emerging learning algorithms that are gaining a lot of attention due to their simplicity and high model generalisation accuracy [233],[236],[18],[19]. This attention is also accredited to the ability of ML-ELMs to estimate in a fast manner the parameters of hidden neurons without a fine tuning [237].

For instance, in [18], an ELM autoencoder was reported. Unlike previous autoencoders where the weights between the hidden and the output layer are randomly selected, the authors introduced an optimisation method to improve this

selection to improve the accuracy and the speed of the neural network. Classification tests with one-dimensional databases showed that the method reached percentages higher than 94% accuracy.

In [7], an ELM-based method for extracting features from benign and malignant ROIs was presented to increase the convergence speed of training and achieve better generalisation properties. However, extraction efficiency was highly dependent on another method to select the feature values. Testing on a broad set of mammograms (949) and 12 classes showed that the method reached an accuracy of 91%. In [206], a deep autoencoder based on multiobjective optimization was implemented on the one hand to reduce the dimensionality of the data to be useful in classification and on the other to reduce the *reconstruction error* existing between the first (encoder input) and the last layer (decoder output) of the autoencoder to reduce the classification error. Tests with different classifiers in the last layer of the network gave accuracies ranging between 80% and 98%.

This chapter reports a Multilayer Fuzzy Extreme Learning Machine (ML-FELM) based on the practical equivalence between FLSs and the Multi-Input-Multi-Output RBFNN model for breast cancer image classification. The ML-FELM follows the hierarchical learning process of an ML-ELM [19] and the ML Kernel-ELM [20]. In other words, the parameter identification of the ML-FELM consists of two main steps. At first step, some stacked Fuzzy Autoencoders (FAEs) are used as a mechanism for data representation by retrieving a set of high-quality features and then classified in a second step by using a MIMO IT2-RBFNN. Unlike the ML-KELM, the proposed ML-FELM computes the normalised weighted average, which is a defuzzification mechanism at each FAE's output layer. Therefore, each FAE in the first step is a MIMO RBFNN that is practically equivalent to a

MIMO FLS where the process of fuzzification, inference engine and the defuzzification is carried out in the hidden and output layer respectively.

Two separate experiments are performed to study the effectiveness of the proposed ML-FELM. First, to compare the computational time that is required to train the ML-FELM for big data, the MNIST dataset for handwritten digits is employed. Similar to ML-ELM [19] and ML-KELM [20], the ML-FELM is much faster for the classification of large data sets than other deep learning structures such as the DBN and DBM [20]. Finally, the mini-MIAS breast cancer image repository is employed to evaluate the ability of the ML-FELM for data representation and image classification. According to our results, for breast cancer classification the ML-FELM outperforms other machine learning methodologies such as the OCI-ELM [18], ELM [7], GPZM [204], 2D-NARX [229] and similar to the SVM-ELM [205], MOEA-2c [206].

The structure of the rest of the chapter is the following: in Section 5.2 a short review of Extreme Learning Machine and its application to the Radial Basis Function Neural Network (RBFNN) is provided, as well as a description about the practical equivalence between Fuzzy Logic Systems (FLSs) and the RBFNN of either Type-1 or Interval Type-2. Section 5.3 describes the proposed ML-FELM, while Section 5.4 provides experiments and evaluation of the ML-FELM. Finally, concluding remarks are provided in Section 5.5.

5.2 Preliminaries and definitions

5.2.1 ELM and multi-input-multi-output RBFNN

ELM was initially proposed as a single hidden layer feedforward network (SLFNN) [237] to easily achieve a high generalisation performance at super fast speed. Then ELM was extended to other neural structures in the form of Radial Basis Function Neural Network (RBFNN)[238]. The jth output of an RBFNN with M kernels and a number of P arbitrary samples $(\mathbf{x}_p, t_p)$ is given by:

$$y_j = \sum_{i=1}^{M} \beta_i g_i(\mu_i, \sigma_i, \mathbf{x}_p),\ j = 1, \cdots, \tilde{N} \tag{5.1}$$

where $g_i = f_i / \sum_{i=1}^{M} f_i$ are the normalised basis functions, $\mathbf{x}_p = [x_{k_1}, \cdots, x_{k_N}]^T \in \mathbf{R}^N$ is the input vector, $t_p = [t_{p_1}, \cdots, t_{p_{\tilde{N}}}] \in \mathbf{R}^{\tilde{N}}$ is the desired pattern and $\beta_i = [\beta_{i1}, \cdots, \beta_{i\tilde{N}}]^T$ is the vector of weights linking the ith kernel to the jth output. The normalised basis functions g_i are given by:

$$g_i(\mu_i, \sigma_i, x_p) = \exp\left[-\sum_{k=1}^{N} \left(\frac{x_k - \mu_{k_i}}{2\sigma_i} \right)^2 \right], i = 1, \cdots, M \tag{5.2}$$

As stated in [237], the standard Gaussian RBF kernels can obtain P arbitrary samples with error means equal to zero, that is, given some random parameters μ_{ki} and σ_i, the training of the RBFNN is practically a least-squares solution of a compact linear system. Their formulation is:

$$\mathbf{H}\beta = \mathbf{T} \tag{5.3}$$

in which $\mathbf{H}$ is the output matrix of the hidden layer of the RBFNN concerning the input-output vectors $(\mathbf{x}_p, y_p)$. More formally:

$$\mathbf{H}(\mu_1, \cdots, \mu_M, \sigma_1, \cdots, \sigma_M, \mathbf{x}_1, \cdots, \mathbf{x}_P) = \begin{pmatrix} f_1(\mu_1, \sigma_1, \mathbf{x}_1) & \cdots & f_M(\mu_M, \sigma_M, \mathbf{x}_1) \\ \vdots & \vdots & \vdots \\ f_1(\mu_1, \sigma_1, \mathbf{x}_P) & \cdots & f_M(\mu_M, \sigma_M, \mathbf{x}_P) \end{pmatrix}_{P \times M}$$

where the vector of centres is $\mu_i = (\mu_{1_i}, \cdots, \mu_{N_i})$ and the vectors of weights and outputs are,

$$\beta = \begin{pmatrix} \beta_1^T \\ \vdots \\ \beta_M^T \end{pmatrix}_{M \times \tilde{N}} \quad \text{and} \quad \mathbf{T} = \begin{pmatrix} y_1^T \\ \vdots \\ y_P^T \end{pmatrix}_{P \times \tilde{N}} \tag{5.4}$$

The solution of the linear system $\mathbf{H}\beta = T$ by the minimum norm least-squares is unique and may be obtained by estimating the pseudoinverse $H^\dagger$ so that:

$$\beta = \mathbf{H}^\dagger T \tag{5.5}$$

In line with the Ridge theory of regression, a positive value specified by the user $1/\mathbf{C}$ can be added to have a better generalisation performance [154].

$$\beta = \left(\frac{\mathbf{I}}{\mathbf{C}} + \mathbf{H}^T \mathbf{H} \right)^{-1} \mathbf{H}^T \mathbf{T} \tag{5.6}$$

5.2.2 MIMO IT2-RBFNN and fuzzy logic

It has been shown that in ideal conditions, an RBF network may be seen either as a Type-1 (T1) or as Interval Type-2 Fuzzy Logic System (IT2 FLS) [239],[240]. This equivalence is used in a number of applications for the modelling of complex systems [240],[241] [242]. Figure 5.1 depicts the interaction of fuzzy components during an inference process. In it, the inference engine uses the rule base to compare the input image values to the outputs to generate an interval Gaussian membership function to be defuzified later in the network.

Generally speaking, an RBFNN can be considered an FLS of T1 or IT2 (for short IT2-RBFNN) when its neural structure consists of [239],[240],[241]:

1. The input layer with singleton fuzzification and Membership Functions (MFs) within the rules are taken as Gaussian neurons [243].

2. The operator T-norm employed to compute the rule firing strengths in the hidden layer is a minimum.

3. The secondary membership function of each fuzzy system is convex and either of T1 or IT2.

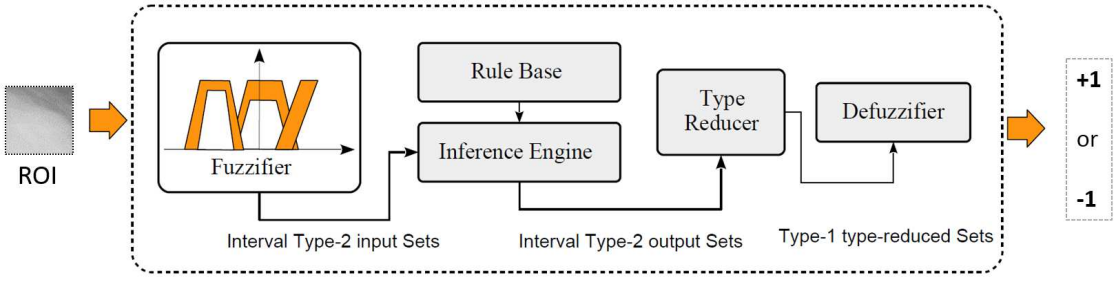


Figure 5.1: Fuzzy logic process implicit in the neural network (ROI from [212]).

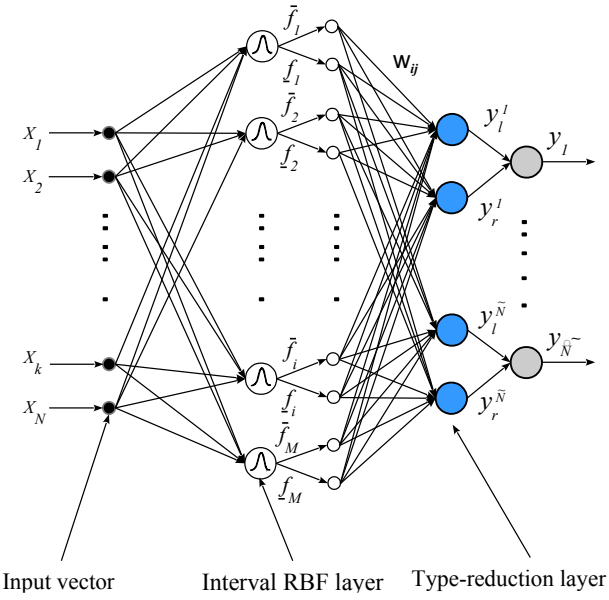


Figure 5.2: Multiple-input-multiple-output Interval Type-2 RBFNN

4. The RBF and the fuzzy inference system employ the same procedure (either weighted average or summation) to obtain each global output. The defuzzification formula mainly depends on the order of the FLS.

More specifically, every enhancement made by FLS can be useful to the RBF theory since the its fuzzy rule structure base goes from T1 Fuzzy Sets (FSs) to higher order FSs remains the same; in this form the modelling of the linked antecedents and consequents takes place [244]. In that way, to design an RBFNN that is practically equivalent to a Multi-Input-Multi-Output (MIMO) IT2 FLS with a Karnik-Mendel algorithm, the associated inference mechanism must be interpreted as an adaptive filter which resembles an additive combination of the MFs (firing strengths). As illustrated in Figure 5.2, each associated fuzzy rule R^i in a MIMO IT2-RBFNN is described by a multi-variable Gaussian MF

$\mu_{R^i}(\mathbf{x}_p, y_p^j) = \mu_{R^i}[x_1, \cdots, x_n, y_p^j]$, where the input vector $\mathbf{x}_p \in X_1 \times \cdots X_n$ and the inference engine is defined as:

$$\mu_{R^i}(\mathbf{x}_p, \beta) = \mu_{A^i \to \beta} = \left[T_{k_1}^N \mu_{F_k^i}(x_k) \star \mu_{G^i(\beta)} \right] = [\underline{f}_i(\vec{x}_p), \overline{f}_i(\vec{x}_p)] \tag{5.7}$$

in which $\star$ represents the $t-norm$ minimum, or the smaller distance to the vector of inputs $\mathbf{x}_p$, and $[\underline{f}_i(\vec{x}_p), \overline{f}_i(\vec{x}_p)]$ are the lower and upper membership functions (LMF, UMF) respectively. In this chapter, each MF in the MIMO IT2-RBFNN fuzzy rule is an interval Gaussian MF with an uncertain width $\sigma_i = [\sigma_i^1, \sigma_i^2]$ and fixed center (mean) μ_{ki}, as it is shown in Figure 5.3.

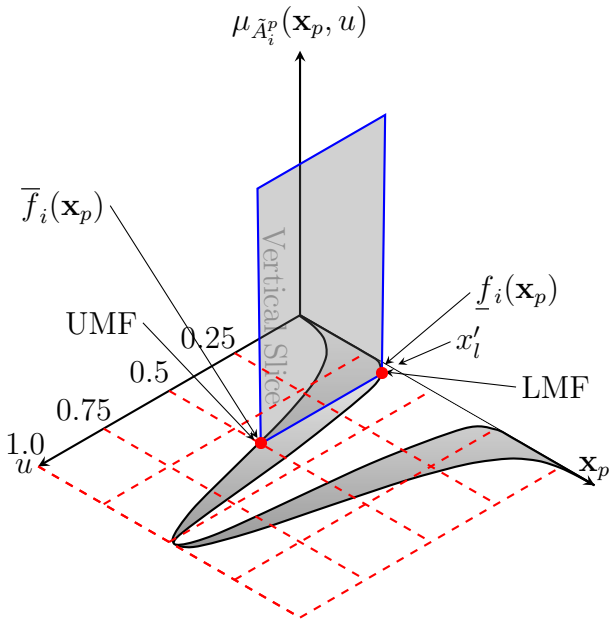


Figure 5.3: Singleton fuzzification and interval secondary MF which becomes active if $\mathbf{x}_p = x_l'$ for the ith recipient unit of the network (taken from [245]).

$$F^i := \begin{cases} F^i = [\underline{f}_i(\mathbf{x}_p), \overline{f}_i(\mathbf{x}_p)] \\ \underline{f}_i(\mathbf{x}_p) = \exp\left[-\sum_{k=1}^{N}\left(\frac{x_k - \mu_{k_i}}{2\sigma_i^2}\right)^2\right] \\ \overline{f}_i(\mathbf{x}_p) = \exp\left[-\sum_{k=1}^{N}\left(\frac{x_k - \mu_{k_i}}{2\sigma_i^1}\right)^2\right] \end{cases} \tag{5.8}$$

The ith fuzzy rule of a MIMO IT2-RBFNN is stated as follows:

$$\tilde{R}^i : \text{IF } x_1 \text{ is } F_1^i \text{ and} \cdots \text{IF } x_k \text{ is } F_s^i \text{ and} \cdots \text{IF } x_N \text{ is } F_N^i \text{ THEN } y \text{ is } w_{ij};\ i = 1, \cdots, M \tag{5.9}$$

In the case of a Mamdani type IT2-RBFNN (also known Zadeh type), the weight vector (consequent) w_i is a vector of single crisp (non-fuzzy) values, while for a TSK model, each $w_i = c_0^i + c_1^i x_1 + c_2^i x_2 + \cdots + c_N^i x_N$. By considering the practical equivalence among the RBFNN and the IT2 FLSs [240], for each output y_j, the MIMO IT2-RBFNN is a FLS with a reduction of the type centre-of-sets, rule of product inference, and the output space of a singleton. The reduction-type set (y_l, y_r) portrayed earlier in Figure 5.2, results from a Karnik-Mendel algorithm [246]. In agreement with figures 5.1 and 5.2, when w_{ij} is a crisp value and the inference engine for the IT2-RBFNN can be of Mamdani or TSK type, the formulation of the matrix for the jth output in the MIMO IT2-RBFNN is stated as [244],[246]:

$$y_j = \frac{1}{2}\left(\mathbf{Y}_l^j + \mathbf{Y}_r^j\right)\mathbf{w}_{ij}^T \tag{5.10}$$

in which the outputs $y_l^j = \mathbf{Y}_l^j \mathbf{w}_{ij}^T$ and $y_r^j = \mathbf{Y}_r^j \mathbf{w}_{ij}^T$, and where:

$$\mathbf{Y}_l^j = \frac{\bar{\mathbf{f}}^T Q_j^T E_{1j}^T E_{1j} Q_j + \underline{\mathbf{f}}^T Q_j^T E_{2j}^T E_{2j} Q_j}{r_l^T Q_j \bar{\mathbf{f}} + s_l^T Q_j \underline{\mathbf{f}}} \tag{5.11}$$

with $\mathbf{Y}_l^j = (\psi_{lj,1}, \cdots, \psi_{lj,M})$, and the terms E_{1j}, E_{2j}, r_{lj} and s_{lj} are defined as:

$$\begin{aligned}
E_{1j} &= (e_{1j}|e_{2j}|\ldots|e_{Lj}|\mathbf{0}|\ldots|\mathbf{0})^T \;\; L_j \times M \\
E_{2j} &= \left(\mathbf{0}|\ldots|\mathbf{0}|\xi_1^j|\xi_2^j|\ldots|\xi_{M-L_j}^j\right)^T \;\; (M - L_j) \times 1 \\
r_{lj} &\equiv (\underbrace{1, 1, \ldots, 1}_{L}, 0, \ldots, \ldots, 0)^T \;\; M \times 1 \\
s_{lj} &\equiv (0, \ldots, \ldots, 0 \overbrace{1, 1, \ldots, 1}^{M-L_j})^T \;\; M \times 1
\end{aligned}$$

where L_j is the switching point that corresponds to the jth output, $e_m \in R_j^L$ $(m = 1, \cdots, L_j)$ and $\xi_m \in R^{M-L_j},\ m = 1, \cdots, M - L_j$ are the basic vectors in which the values equal zero excepting the mth one, which becomes 1.

$$\mathbf{Y}_r^j = \frac{\underline{\mathbf{f}}^T Q_j^T E_{3j}^T E_{3j} Q_j + \bar{\mathbf{f}}^T Q_j^T E_{4j}^T E_{4j} Q_j}{r_r^T Q_j \underline{\mathbf{f}} + s_r^T Q_j \bar{\mathbf{f}}} \tag{5.12}$$

where $\mathbf{Y}_r^j = (\psi_{rj,1}, \cdots, \psi_{rj,M})$

$$\begin{aligned}
E_{3j} &= (e_{1j}|e_{2j}|\ldots|e_{Rj}|\mathbf{0}|\ldots|\mathbf{0})^T \;\; R_j \times M \\
E_{4j} &= \left(\mathbf{0}|\ldots|\mathbf{0}|\xi_{1j}|\xi_{2j}|\ldots|\xi_{M-R_j}\right)^T \;\; (M - R_j) \times 1 \\
r_{rj} &\equiv (\underbrace{1, 1, \ldots, 1}_{R_j}, 0, \ldots, \ldots, 0)^T \;\; M \times 1
\end{aligned}$$

$$s_{rj} \equiv (0, \ldots, \ldots, 0 \overbrace{1, 1, \ldots, 1}^{M-R_j})^T \; M \times 1$$

where $e_m \in R^{Rj}$ $(m = 1, \cdots, R_j)$ and $\xi_m \in R^{M-R_j}$, $j = 1, \cdots, M - R_j$ represent the basic vectors where the elements equal zero apart from the jth one that becomes 1 [247]. $\underline{\mathbf{f}} = (\underline{f}_1, \cdots, \underline{f}_M)^T$, $\overline{\mathbf{f}} = \left(\overline{f}_1, \cdots, \overline{f}_M\right)^T$. When the Karnik-Mendel algorithms are used [246], the reordered consequent weight $\tilde{\mathbf{w}}_j$ derived from the permutation to finding the switching points L and R is obtained as follows [247]:

$$\tilde{\mathbf{w}}_j = Q_j \mathbf{w}_j^T, \quad Q_j \in R^{M \times M} \tag{5.13}$$

where $\mathbf{w}_j = (w_{1j}, \cdots, w_{Mj})$ represents the set of initial consequent weights ordered by rule and Q_j is the correspondent permutation matrix [247]. Therefore, the defuzzified pth MIMO IT2-RBFNN output is the vector of $\tilde{N}$ outputs $\mathbf{Y}_p = [y_1, \cdots, y_{\tilde{N}}]^T$.

5.2.3 MULTILAYER KERNEL EXTREME LEARNING

The Multilayer Kernel Extreme Learning machines (ML-KELM) are kernel-based multilayer networks that adopt two separate learning steps. As indicated in [20] and portrayed in Figure 5.4, in the beginning a set of high-quality features is obtained using a number of L stacked kernel-based Autoencoders (AEs) where each AE learns to transform data from its hidden to its output layer. At layer zero (the input layer), an input matrix $\mathbf{X}^n$ is mapped into a kernel space $\Omega^{(n)}$

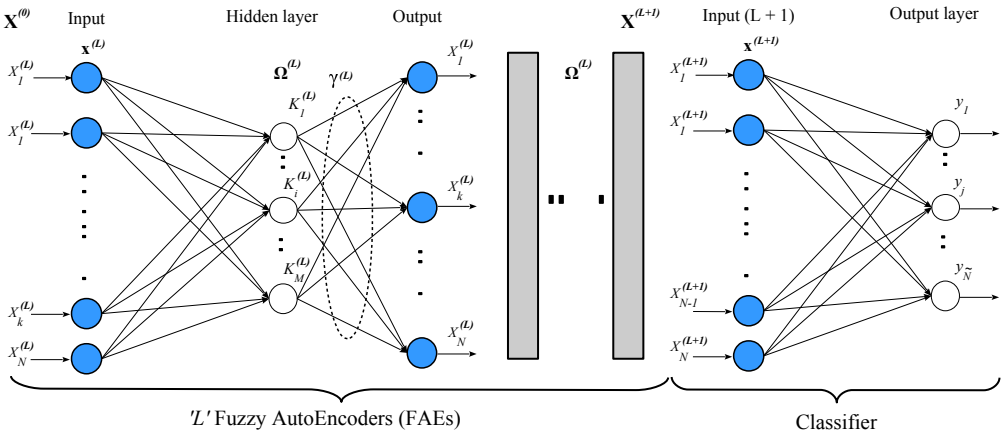


Figure 5.4: Architecture of a Multilayer Kernel Extreme Learning Machine (from [20]).

via the Gaussian activation function $K^{(n)} = \exp(-||(\mathbf{x}_p - \mathbf{x}_q)||/2\sigma^2)$, where a transformation matrix $\Gamma^{(n)}$ is estimated as follows:

$$\Omega^{(n)}\tilde{\Gamma}^{(n)} = \mathbf{X}^{(n)} \tag{5.14}$$

where n is used to indicate the nth data transformation, $\Omega = \mathbf{H}\mathbf{H}^T$, and $\tilde{\Gamma}^{(n)}$ is calculated as:

$$\tilde{\Gamma}^{(n)} = \left(\frac{\mathbf{I}}{\mathbf{C}} + \Omega^{(L)}\right)^{-1} \mathbf{X}^{(n)} \tag{5.15}$$

such as $\tilde{\Gamma}^{(n)} = [\gamma_1^{(n)}, \cdots, \gamma_M^{(n)}]$, where $\gamma^{(n)}$ is the nth transformation vector employed in the learning of the representation of the input data $\mathbf{X}^{(n)}$. The final transformation $\mathbf{X}^{(n+1)}$ is obtained using a sigmoid:

$$\mathbf{X}^{(n+1)} = g\left(\mathbf{X}^{(n)}(\tilde{\Gamma}^{(n)})^T\right) \tag{5.16}$$

As pointed out in [154], [20],[236], if the nth transformation holds identical dimension that the $(n+1)$th layer, the activation function g may be selected as linear piecewise.

5.3 Multilayer fuzzy extreme learning machine

The proposed Multilayer Fuzzy Extreme Learning Machine (ML-FELM) is a multilayer fuzzy network inspired on the practical correspondence among the FLSs and the RBF. Similarly to ELM for a multilayer perceptron [19], ML-FELM has a multilayer structure as displayed in Figure 5.4. Structurally speaking, the proposed ML-FELM is composed of $N+1$ layers whose parameter identification consists of two main steps, that is, (a) an unsupervised hierarchical feature representation stage [154],[19], [248], and (b) a supervised feature classification [240],[241].

The first step consists of a process for high-quality feature extraction by stacking L Fuzzy Autoencoders. In other words, the nth layer uses an independent MIMO T1 RBFNN (which is practically correspondent to a type 1 FS) as a Fuzzy Autoencoder (FA). Unlike the ML-ELM suggested in [154] and [19], at the input layer (layer *zero*) the input data does not need to be converted into an ELM space of random features. Therefore, in the first step, representation learning of the input data is initially performed, as the following equation shows:

$$\mathbf{H}^{(n)}\Gamma^{(n)} = \mathbf{X}^{(n)},\ n = 1, \cdots, L \tag{5.17}$$

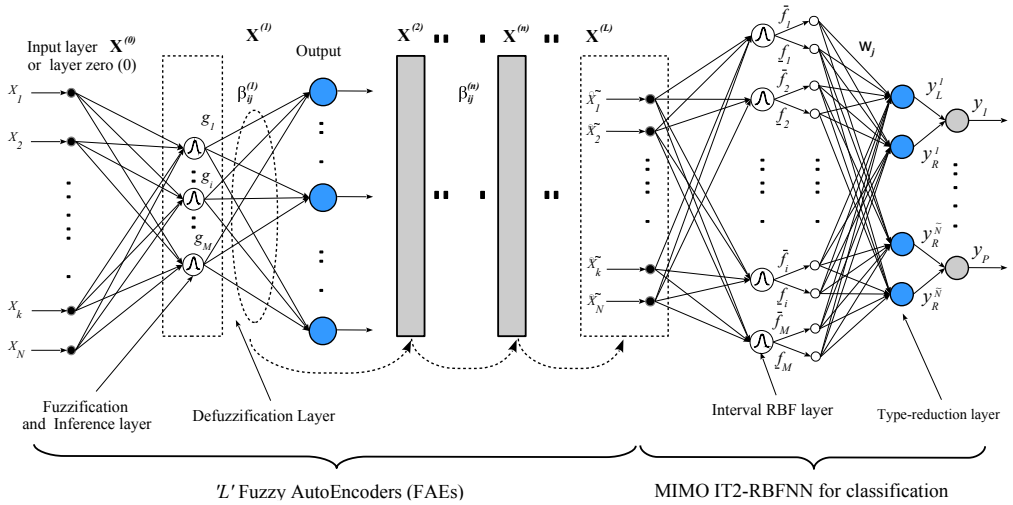


Figure 5.5: Fuzzy Autoencoder and MIMO IT2-RBFNN based on Extreme Learning Machines.

in which, $\mathbf{H}^{(n)} = [h(\mathbf{x}_1), \cdots, h(\mathbf{x}_P)]$, $p = 1, \cdots, P$, and the vector $h(\mathbf{x}_p) = [A_1(\mathbf{x}_p), \cdots, A_M(\mathbf{x}_p)]$, where the term A_i is the ith normalised firing strength that is calculated using the equation below (see Figure 5.5):

$$A_i(\mu_i, \sigma_i, \mathbf{x}_p) = \frac{f_i(\mu_i, \sigma_i, \mathbf{x}_p)}{\sum_{i=1}^{M} f_i(\mu_i, \sigma_i, \mathbf{x}_p)} \tag{5.18}$$

where the parameters (μ_i, σ_i) of each Gaussian MF in the nth layer are randomly selected. Hence, each output y_j in the FAE is a weighted average that plays the role of a defuzzification process in T1 FLSs as follows:

$$y_j = \sum_{i=1}^{M} A_i(\mu_i, \sigma_i, \mathbf{x}_p)\beta_{ij}, \;\; j = 1, \cdots, \tilde{N} \tag{5.19}$$

in which the pth FAE output $\mathbf{y}_p = [y_1, \cdots, y_{\tilde{N}}]^T$ is used to build $\mathbf{X}^{(n)}$. Therefore, the transformation matrix $\tilde{\Gamma}^{(n)}$ is computed as follows [20]:

$$\tilde{\Gamma}^n = \left(\frac{\mathbf{I}}{C} + \mathbf{H}^{(n)}(\mathbf{H}^{(n)})^T\right)^{-1} (\mathbf{H}^{(n)})^T \mathbf{X}^{(n)} \tag{5.20}$$

In this way, $\mathbf{X}^{(n+1)}$ can be computed using (5.16). In a similar way to [154], $g(\cdot)$ can be any activation function. However, if the dimension of layer (n) and layer $(n+1)$ is the same, it is recommended to chose a linear piecewise function [20]. In the second step, the layer $L+1$ is an IT2-RBFNN whose inputs are the resultant high-quality features obtained at layer L by the latest FAE.

Therefore, to find the parameters of the MIMO IT2-RBFNN, ELM is systematically used in two separate moments to refresh the corresponding weights in the IT2-RBFNN output layer [245, 249]. The first step [249] includes the acquisition of the optimal initial values for the corresponding ones by estimating the reduction-type set for the jth output $[y_l^j, y_r^j]$ as:

$$y_{l,1}^j = \frac{\sum_{i=1}^M \underline{f}_i w_{ij}}{\sum_{i=1}^M \underline{f}_i} = \sum_{i=1}^M \underline{f}'_i w_{ij};\ \underline{f}'_i = \frac{\underline{f}_i}{\sum_{i=1}^M \underline{f}_i} \tag{5.21}$$

$$y_{r,1}^j = \frac{\sum_{i=1}^M \overline{f}_i w_{ij}}{\sum_{i=1}^M \overline{f}_i} = \sum_{i=1}^M \overline{f}'_i w_{ij};\ \overline{f}'_i = \frac{\overline{f}_i}{\sum_{i=1}^M \overline{f}_i} \tag{5.22}$$

where the weight vector $\mathbf{w}_j = [w_{1_j}, \cdots, w_{M_j}]^T$ and the weight matrix is stated as $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{\tilde{N}}]$. By using Equations 5.11 and 5.12, the following linear system can be defined for P patterns as follows:

$$\mathbf{T} = \Phi_{\mathbf{A}}(\mathbf{X}^{(L+1)})\mathbf{W}, \mathbf{W} \in \mathbf{R}^{M \times \tilde{N}} \tag{5.23}$$

where $\mathbf{T} = [t_1, \cdots, t_P]^T$, is the desired output vector, $p = 1, \cdots .P$ and each $t_p = [t_{1_P}, \cdots, t_{M_p}]^T$. The matrix Φ_A is defined for an IT2-RBFNN having a Mamdani fuzzy rule structure as:

$$\Phi_{\mathbf{A}}(\mathbf{x}) = \begin{pmatrix} \Phi_1 \\ \Phi_2 \\ \vdots \\ \Phi_P \end{pmatrix} \in R^{P\times(M\times N)} \tag{5.24}$$

From Equation 5.21 and 5.22 it follows for a TSK implication:

$$\Phi_p \mathbf{W}_j = \frac{1}{2}\sum_{i=1}^{M}(\overline{f}'_i + \underline{f}'_i)\left[\sum_{k=1}^{N} c_k^{i_1} x_k, \cdots, \sum_{k=1}^{N} c_k^{i_{\tilde{N}}} x_k\right];\ j = 1, \cdots, \tilde{N} \tag{5.25}$$

For a Mamdani type IT2-RBFNN, the second sum term in Equation 5.25 is a crisp number w_{ij}. In that way, the system described in Equation 5.23 can be solved as below:

$$\mathbf{W}_1 = \Phi_{\mathbf{A}}(\mathbf{x})^{\dagger}\mathbf{T} \tag{5.26}$$

where $\mathbf{W}_1$ is the optimal primary value for the correspondent matrix $\mathbf{W}$ and $\Phi_{\mathbf{A}}(\mathbf{x})^{\dagger}$ is the Moore-Penrose generalised inverse of $\Phi_{\mathbf{A}}(\mathbf{x})$. Then, the last optimisation process of $\mathbf{W}$ consists of implementing jth times the Karnik-Mendel algorithm. In other words, each column vector in the matrix $\mathbf{W} = [\mathbf{w}_1, \cdots, \mathbf{w}_{\tilde{N}}]^T$ is used to form a linear system given by:

$$\mathbf{t}_j = \Phi_{\mathbf{B}}^{j}(\mathbf{x})\mathbf{w}_j, \mathbf{w}_j \in \mathbf{R}^M \tag{5.27}$$

Therefore, from equations 5.11 and 5.12 the terms $\mathbf{Y}_l^j$ and $\mathbf{Y}_r^j$ are used to calculate each weight vector $\mathbf{w}_j$ where each $\Phi_{\mathbf{B}}^j(\mathbf{x})$ becomes:

$$\Phi_{\mathbf{B}}^j(\mathbf{x}) = \begin{pmatrix} \tilde{\Phi}_1 \\ \tilde{\Phi}_2 \\ \vdots \\ \tilde{\Phi}_P \end{pmatrix} \in R^{P\times(M\times N)} \tag{5.28}$$

so that:

$$\tilde{\Phi}_p \mathbf{w}_j = \frac{1}{2}\sum_{i=1}^{M}(\psi_{l,i} + \psi_{r,i}) \left[\sum_{k=1}^{N} c_k^{i1} x_k, \cdots, \sum_{k=1}^{N} c_k^{i\tilde{N}} x_k\right] \tag{5.29}$$

5.4 EXPERIMENTS AND RESULTS

The evaluation of the performance of the proposed ML-FELM networks included two different experiments using two different image data sets. First, the MNIST data set was used to compared the ML-FELM performance against other existing techniques. Secondly, the ML-FELM was applied to breast cancer image classification using the mini-MIAS data repository. The experiments carried out were processed with Intel (R) Core (TM) i5-4590 at 3.30GHz speed, running MATLAB 2014b. The evaluation of the effectiveness of the MIMO IT2-RBFNN required the implementation of an ML-FELM having a MIMO RBFNN in the $L+1$ layer (ML-FELM-RBFNN for short). Then, this section denotes the short name ML-FELM-IT2-RBFNN to describe an ML-FELM with a MIMO IT2-RBFNN in the $L+1$ layer.

5.4.1 Classification of handwritten digits

As mentioned at the beginning of the chapter, the initial testing stage required, before the cancer detection tests, the use of a handy, ready-to-use and comparable database to train and evaluate the new model multiple times. It was fundamental to know if the feature extraction process was delivering consistent results to be effectively interpreted by the classifier. Hence the first data set used in this chapter was the MNIST database [250], which consists of a collection of 70,000 images of 28×28 pixels containing handwritten digits with subsets of 60,000 and 10,000 images for training and testing, respectively.

The MNIST database is a well-known data set to test deep neural structures and machine learning algorithms. Analogously to the results presented in [154], in this chapter the MNIST is used without any distortion to compare the performance of the suggested ML-FELM. Table 5.1 details a performance comparison between the ML-FELM, a Deep Belief Network (DBN), a Deep Boltzmann Machine (DBM), a Stacked AutoEncoder (SAE) and a Deep Network based on Orthogonal convex incremental ELM (OCI-ELM) [18].

To keep a good balance between model precision and low computational load, the experiment setup for the ML-FELM-IT2-RBFNN and ML-FELM-RBFNN with $L = 3$ is $C = [0.1, 50^{-5}, 10^{4}, 10^{8}]$ for 784-300-300-1000-10 hidden neurons. The hidden neurons quantity for the DBN, DBM and the ML-ELM are 784-500-500-2000-10, 784-500-1000-10 and 784-700-700-15000-10 respectively. As can be noted from Table 5.1, the improvements offered by the proposed ML-FELM do not make it the fastest learner. However, an important increase in model accuracy and a simpler neural structure compensates this limitation. Moreover, by adding

Table 5.1: Comparison between the new ML-FELM and previous ML networks in the MNIST database.

Model	Reference	Accuracy %	Trainig Time
ML-ELM	[154]	99.03 (± 0.04)	444.655s
ELM (random features)	[20]	97.31 (± 0.1)	545.95s
ELM (Gaussian kernels)	[20]	98.75	790.96s
DBN	[20]	98.87	20580s
DBM	[20]	99.05	68246s
OCI-ELM	[18]	97.94	3985s
SAE	[251]	98.6	-
ML-FELM-RBFNN	-	98.57 (± 0.13)	2610.9s
ML-FELM-IT2-RBFNN	-	99.14 (± 0.08)	2870.1s

an IT2-RBFNN in the $L+1$ layer, the generalisation properties of a MIMO RBFNN are significantly improved.

5.4.2 Breast cancer classification and detection

The second data repository for the evaluation of the ML-FELM-RBF and ML-FELM-IT2-RBF neural networks is the mini-MIAS database of mammograms, a public image data source containing a selection of 322 digital mammograms of 1024×1024 pixels, gathered from the United Kingdom National Breast Screening Program [212]. The repository contains radiological images captured from the mediolateral oblique view of the breast representing incidences of common abnormalities (66 benign and 52 malignant) and normal or healthy cases. Extensive documentation of the image database includes location, diameter and class of abnormalities.

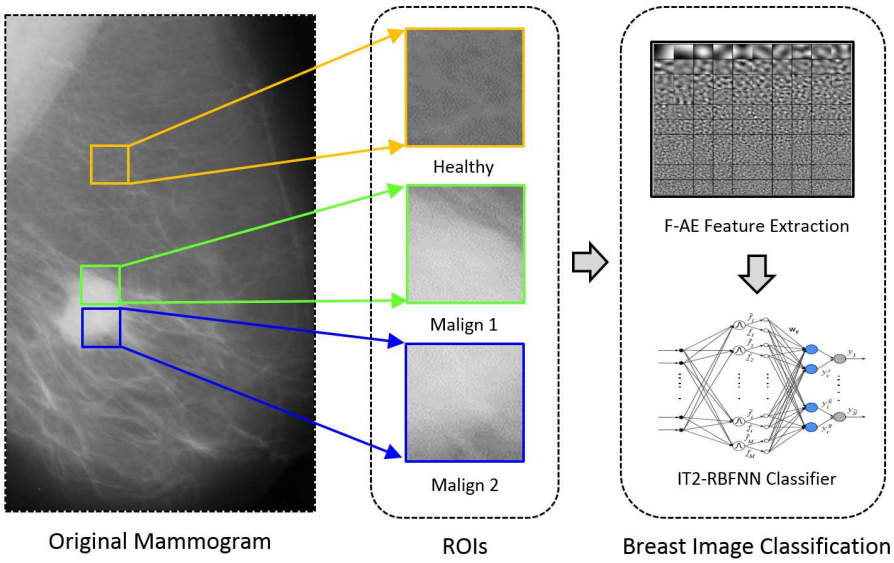


Figure 5.6: Fuzzy Autoencoder and MIMO IT2-RBFNN based on Extreme Learning Machines (mammogram mdb028, from [212]).

Due to the size of the mini-MIAS data set and the number of labels and annotations in each image, an initial preprocessing step aimed at creating a broader set of subimages or Regions of Interest (ROI) was carried out. In other words, it was necessary to create sets formed by many ROIs containing abnormalities in the case of benign and malignant samples or healthy tissue for ordinary cases. Thus, as illustrated in Figure 5.6, an additional procedure was performed to expand the number of benign and malignant samples while improving the variability of the training set. To enhance the variability prediction of the ML-FELM, the final number of ROIs includes a subset of rotated images at 90, 180 and 270 degrees.

The increasing of samples mentioned above produced in total 4200 samples. Such number came from 350 original samples, three uniformly distributed classes (benign, malignant and healthy) and 4 rotation angles, totalling $350 \times 3 \times 4 = 4200$ ROIs. The experiment setup for the ML-FELM using a MIMO IT2-RBFNN and RBFNN is with a neural structure of 3 layers, where $C = [0.16, 0.004, 10^6, 10^8]$, and

with many hidden neurons 4096-1500-1500-4800-2. The dataset was split into two subsets for cross-validation purposes, that is, 65% and 35% of the total of images are used for training and testing, respectively. To measure the associated model performance, the metrics accuracy, sensitivity and specificity, detailed in Section 2.4.2, are used in this chapter. A comparison between the proposed ML-FELM and other existing methodologies for the diagnosis of breast cancer is presented in Table 5.2.

The comparison table shows a similar classification performance in most cases, although the feature extraction stage offered and the number of training samples is different in each methodology. In particular, MOEA-2c, OCI-ELM and ML-FELM show the highest accuracy performance. It is worth mentioning, the

Table 5.2: Performance comparison between the proposed ML-FELM and previous machine learning techniques for breast cancer detection.

Model	**Ref.**	**Image Set**	**Acc. %**	**Sen. %**	**Spe. %**
MOEA-2c	[206]	Inbreast+IRMA	93.4-98.1	-	-
OCI-ELM	[18]	UCI: Br. Canc.	94.73	-	-
SVM-ELM	[205]	DDSM	95.73	94.8	97.16
ELM	[7]	mini-MIAS	91.00	90.00	98.00
GPZM	[204]	mini-MIAS	89.38	83.58	93.43
GPZM	[204]	DDSM	87.27	82.51	90.33
2D-NARX	[217]	mini-MIAS	91.00	93.00	89.50
MSRBF DCT	[229]	mini-MIAS	93.50	87.00	96.90
ML-FELM-RBF	-	mini-MIAS	92.65	94.31	90.94
ML-FELM-IT2-RBF	-	mini-MIAS	95.13	94.14	97.78

implementation of a MIMO IT2-RBFNN in the $L+1$ layer produces a better generalisation performance compared to its T1 fuzzy counterpart.

5.5 Discussion

- This chapter reports a Fuzzy Multilayer Fuzzy Extreme Learning Machine (ML-FELM) considering the practical equivalence among the RBFNN and Fuzzy Logic Systems (FLSs).
- Similarly to the ML-ELM and the ML-KELM, the parameter identification of the ML-FELM consists of two phases. The first step consists of many stacked Fuzzy Autoencoders (FAEs) that are practically equivalent to FLSs of type-1 for the extraction of high-quality features. Consequently, the second step contemplates a MIMO IT2-RBFNN that acts under the Interval Type-2 FLS logic for classification purposes of the encoded data.
- In the same way to other existing ELM approaches, the proposed method offers the following advantages:

    1. The ML-FELM does not need fine tuning.
    2. The implementation of an IT2-RBFNN enhances the generalisation properties of the RBFNN for data classification.
    3. Compared to ML-ELM, an ML-FELM does not need an initial orthogonal feature representation.
    4. According to the results, the computational burden for the parameter identification of the ML-FELM is similar to other ELMs.

- In other words, the proposed ML-FELM is an ML architecture that under ideal circumstances can be seen not only as an FLS but also as a one-forward-

step machine learning able to find a good balance regarding model simplicity, model accuracy and high-quality data representation.

- Regarding the network training time in Table 5.1, it is possible to see that the parameter adjustment of the Gaussian function, which is non-linear, hindered in some way the processing speed. However, the classification runtime per ROI was negligible after the learning completion.

- Finally, it was interesting to find that multiple-output structures, as the network in this chapter, are ideal for classification. Instead, single-output models are better for function approximation, as confirmed in Chapters 3 and 4.

Chapter 6

Conclusions and final considerations

6.1 Summary

The work in this thesis explored the adaptation of non-linear system identification and neural networks (both simple and deep-structured) frameworks into new digital image feature extraction and classification models. The work used the processing of mammographic images as the primary study case, with the intent of classifying the samples according to their clinical condition, for which it also contributed in the exploring and enhancement of computer-aided diagnosis (CAD) systems.

Chapters 1 and 2 presented the problem and recounted the technical and literary background of the solution methods proposed later in the thesis. There was a stress in image processing, system identification and the strong plethora of neural network based learning methods for image classification.

Chapter 3 presented for the first time the NARX models as non-linear image processing methods and as a part of CAD systems for cancer detection. The implementation process took into account the forward regression OLS algorithm as the

solution method for the identification problem. A reinterpretation and modeling of digital images as input-output systems took place to make them compatible with system identification procedures. The motivation of such modelling work obeyed two reasons; (a) that the NARX models are very efficient and easy to identify thanks to the FROLS algorithm and (b) that prior to this work these algorithms had not been used in image processing or feature engineering. As the adaptive capacity of the NARX models prevented the direct creation of equal-sized feature vectors, the design and implementation of a stimulus-response module helped to generate output signals that allowed to obtain a regular number of coefficients from the image model.

Chapter 4 reported, also for the first time, the use of the multiscale radial basis function networks in image processing and as a ground of a new CAD system. The MSRBF combined with the FROLS algorithm had proven to be broadly flexible and efficient in the modelling and solving of nonlinear dynamic systems in difficult study cases. In this study, MSRBF networks turned out to be also very skilled in image modelling. After the model identification, the discrete cosine transform was incorporated to enhance the coefficient extraction of the stimulus-response module. The original FROLS solution algorithm adds candidate terms to the model until the reconstruction error decreases down to an accuracy threshold [88]. In this work, the stop-criterion of FROLS changed to an *IF* function that stopped the algorithm with the accuracy threshold or when the model attained a maximum length. Also, the addition of a two-fold image characterisation to increase the analysis resolution took place in the method.

The capacity of the NARX and MSRBF methods in feature extraction was complemented by the K-means++ classification algorithm to build CAD systems based on the classification of feature vectors derived from images.

Chapter 5 unveiled a new image processing model termed the multilayer fuzzy extreme learning machine (ML-FELM), from which a new CAD system came to light. The ML-FELM is a deep neural network based on stacked autoencoders, radial basis function (kernel-based) neural networks and elements of type-2 fuzzy systems. This last feature sought to deepen the capacity of the classification process by taking into account the uncertainty of the problem concerning the membership degree of the object to different classes.

The new neural network used the first layers to extract feature values from the image through an autoencoder-based ELM. Autoencoders (AEs) are neural networks that learn from the input data in an unsupervised manner to make a more efficient representation. AEs are also able to reduce the data dimensionality as the information moves through the network layers. Kernel-based ELM networks take advantage of such design by making the autoencoder to process data more efficiently via the replacement of dispensable calculations with random values and by removing the back-propagation training process. The last layers of the system (multilayer fuzzy) use the extracted feature values to handle a fuzzy classification which considers overlapping ranges between classes within the decision-making process.

All the contributions were applied as CAD systems using the well-known mini-MIAS public database for the detection of breast cancer as a case study. Before the classification tests, the database was partitioned into subimages and labelled into classes to enable the consistent training of the models.

6.2 Conclusions

The tests from the previous contributions revealed exciting findings. At first, the results showed that the polynomial NARX model, solved through the FROLS algorithm, managed to create precise and adaptive models of the images including non-linear terms while presented particular mathematical structures according to the image composition. The fact of finding non-linear structures in this problem corroborated the existence and proportionality of this critical feature within the surface of digital pictures and enabled a tool for a more in-depth further mathematical-based analysis and quantification. Classification performance values showed that the method was more efficient in detecting abnormalities than to determine normal cases, while by comparing the results with previous work indicated that the technique was more competitive than most models. However, the recurrent use of non-public databases in prior methods hindered a more objective comparison.

The MSRBF neural network, although less transparent in structure than the NARX representation, demonstrated greater flexibility in image modelling. This advantage was proven when none of the thousands of vectors generated by the model was exactly the same, despite processing virtually identical subimages in several cases. The null incidence of identical models for similar images also argues that the two-fold ROI characterisation extended the ability of the model to extract size and object position features from the image successfully. The change of the FROLS stop-criterion into an *IF function* favoured a faster processing of multiple ROIs without sacrificing the modelling efficiency. Experiments done with different breast-type distributions for testing and training revealed that dense mammograms hampered the accuracy of the classifier. In contrast, a direct relationship between

accuracy and fatty type mammograms showed up. Although it was evident that the classification results are susceptible to the type of tissue, little has been found in the experiments reported in the CAD literature that take into account this factor. The overall accuracy of the model was high, although the capacity to detect negative cases was higher than that of positive cases. This difference could be due to an unintended effort to reduce false positives during the manual labelling of the samples.

Tests with the ML-FELM deep neural network showed that it offered a good trade-off between simplicity, feature extraction and classification. The critical elements of the ML-FELM design were (a) the incorporation of fuzzy logic systems to diminish the effects of uncertainty and (b) the autoencoders contained in ELM, that produced a change of dimensionality of the raw input data to make it interpretable by the ML-F classifier, which at first had particular difficulties. For instance, the DCT, a 2D wavelet transform and convolution masks were tried unsuccessfully to code the input data at first. The work in [252] provided a possible explanation by holding that when a network structure is not deep enough (as it is the ML-F network), it may present difficulties at solving non-linearly separable problems. When this is the case, it is necessary to modify the data dimensionality to increase the problem separability. Autoencoders solve this problem thanks to their bottleneck-shaped structure, which forces the data dimensionality to change down. Experiments with two case studies showed that the ML-FELM-IT2 model was superior to most previous methods. In breast cancer detection, it proved to be more effective in detecting negative cases than positive ones, although in both cases the results were encouraging.

As for the case of study, the manual labelling of ROI samples is, despite its disadvantages, necessary in the breast cancer detection problem. This conclusion is

because databases documentation is not always accurate as to the location and radius of tumours, added to the similarity between many healthy dense (or glandular) subimages and ROIs captured from abnormal tissue. Also, the divide and conquer strategy together with the analysis at subimage zoom-level is considered necessary in the problem of breast cancer detection and more specifically in the case of microcalcifications. This type of lesion is in the practice infinitesimal compared to a complete mammogram, while fails to detect it are undoubtedly dangerous, as 78% of this cases stand as extremely suspicious [211].

6.3 Future work

This section lists the opportunity areas and challenges derived from the limitations of this research.

- The feature extraction of the NARX image model structure, which is adaptive and has the potential to provide valuable information in each case. This work did not seize this information because its qualitative and changing nature did not make easy to figure out the way to estimate equal-sized feature vectors.
- A deeper adjustment of the NARX model and the 2D image representation through the receiver operating characteristic (ROC) curve to analyse different paired configurations, including the NARX maximum nonlinear order, the NARX maximum lagged observations, the 2D pixel neighbourhood shape and size, and the step-size between mask image scans.
- The optimal selection of the number of kernel centres of the MSRBF network to reduce the computational load of the FROLS algorithm, given that the estimation code used in this work considers a k value according to the input

data, as it was possible to note that k is suboptimal, and occasionally it does not adapt itself to the problem.

- The exchange between the adopted classifiers (K-means ++ and IT2-RBFNN) and the proposed feature extraction models (NARX-FROLS, MSRBF-FROLS, ML-FELM) with the aim of evaluating their performance more objectively and isolating the agents that contribute the most in both processes (extraction and classification). In this way, to consider the creation of a hybrid method based on the evidence of these tests.

- The extension of experiments with the NARX and ML-FELM methods using low and high ratios of dense and glandular mammograms in the testing set. This could confirm whether the findings found in Chapter 4 on the effect of problematic mammograms on the accuracy of the classification in the breast cancer detection problem are generalizable.

- The strengthening of the ML-FELM network through the fine-tuning of: (a) the user-specified regularisation vector $\mathbf{C}$, and (b) the number of hidden neurons M within the autoencoder. That could give the system a better generalisation performance and more depth to classify images by the refinement of the feature extraction process.

- The use of the ML-FELM network in the difficult object detection problem for the unassisted isolation of ROIs within mammograms, including microcalcifications and the processing of dense problematic images. For instance, the ML-FELM may help to classify the regions within mammograms while a secondary analysis at subimage level could isolate the ROI.

Bibliography

[1] Alan C. Bovik. *The Essential Guide to Image Processing*. Academic Press, 2009.

[2] Melvyn A. Goodale. Transforming vision into action. *Vision research*, 51 (13):1567–1587, 2011.

[3] John Russ. *The image processing handbook*. CRC press, 2016.

[4] Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[5] Constantine Kotropoulos and Ioannis Pitas. *Nonlinear model-based image/video processing and analysis*. John Wiley & Sons, Inc., 2001.

[6] Lúcio F. C. Pessoa and Petros Maragos. MRL-filters: A general class of nonlinear systems and their optimal design for image processing. *IEEE Transactions on Image Processing*, 7(7):966–978, 1998.

[7] Gopalan Vani, Ramaswamy Savitha, and Narasimhan Sundararajan. Classification of abnormalities in digitized mammograms using extreme learning machine. In *2010 11th International Conference on Control Automation Robotics & Vision ICARCV*, pages 2114–2117. IEEE, 2010.

[8] Ioanna. Christoyianni, Athanasios. Koutras, Evangelos. Dermatas, and George. Kokkinakis. Computer aided diagnosis of breast cancer in digitized mammograms. *Computerized Medical Imaging and Graphics*, 26(5):309–319, 2002.

[9] Kulsoom Iftikhar, Shahzad Anwar, Izhar Ul Haq, Muhammad Tahir Khan, and Sayed Riaz Akbar. An Optimal Neural Network Based Classification Technique for Breast Cancer Detection. *Journal of Engineering and Applied Sciences*, 35(1):51–58, 2016.

[10] Jerzy Zielinski, Nidhal Bouaynaya, and Dan Schonfeld. Two-dimensional ARMA modeling for breast cancer detection and classification. In *2010 International Conference on Signal Processing and Communications SPCOM*, pages 1–4. IEEE, 2010.

[11] Jerzy Zielinski and Nidhal Bouaynaya. Statistical sequential analysis for detection of microcalcifications in digital mammograms. In *2010 International Conference on Signal Processing and Communications SPCOM*, pages 1–5. IEEE, 2010.

[12] Hadi Tadayyon, Ali Sadeghi-Naini, Lauren Wirtzfeld, Frances C. Wright, and Gregory Czarnota. Quantitative ultrasound characterization of locally advanced breast cancer by estimation of its scatterer properties. *Medical Physics*, 41(1):012903, 2014.

[13] Jose R. Ayala-Solares, Hua-Liang Wei, Richard Boynton, Simon Walker, and Stephen A. Billings. Modeling and prediction of global magnetic disturbance in near-Earth space: A case study for Kp index using NARX models. *Space Weather*, 14(10):899–916, 2016.

[14] Jose R.. Ayala-Solares, Hua-Liang Wei, and Grant R. Bigg. The variability of the Atlantic meridional circulation since 1980, as hindcast by a data-driven nonlinear systems model. *Acta Geophysica*, pages 1–13, 2018.

[15] Hua-Liang Wei, Stephen Billings, and Michael Balikhin. Prediction of the Dst index using multiresolution wavelet models. *Journal of Geophysical Research: Space Physics*, 109(A7):07212, 2004.

[16] Michael. Balikhin, Otilia. Boaghe, Stephen. Billings, and Hugo Alleyne. Terrestrial magnetosphere as a nonlinear resonator. *Geophysical Research Letters*, 28(6):1123–1126, 2001.

[17] Mofeed Turky Rashid, Mattia Frasca, Abduladhem Abdulkareem Ali, Ramzy Salim Ali, Luigi Fortuna, and Maria Gabriella Xibilia. Nonlinear model identification for Artemia population motion. *Nonlinear Dynamics*, 69(4):2237–2243, 2012.

[18] Chao Wang, Jianhui Wang, and Shusheng Gu. Deep network based on stacked orthogonal convex incremental ELM autoencoders. *Mathematical Problems in Engineering*, Volume 2016 (doi:http://dx.doi.org/10.1155/2016/1649486):1–17, 2016.

[19] Jiexiong Tang, Chenwei Deng, and Guang-Bin Huang. Extreme learning machine for multilayer perceptron. *IEEE Transactions on Neural Networks and Learning Systems*, 27(4):809–821, 2016.

[20] Chi Man Wong, Chi Man Vong, Pak Kin Wong, and Jiuwen Cao. Kernel-based multilayer extreme learning machines for representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 29(3):757–762, 2018.

[21] Lotfi Asker Zadeh. The concept of a linguistic variable and its application to approximate reasoning I. *Information Sciences*, 8(3):199–249, 1975.

[22] Ashutosh Kumar Dubey, Umesh Gupta, and Sonal Jain. Breast Cancer Statistics and Prediction Methodology: A Systematic Review and Analysis. *Asian Pacific Journal of Cancer Prevention*, 16(10):4237–4245, 2015.

[23] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 67(1):7–30, 2017.

[24] Alan Bovik. *Handbook of Image and Video Processing*. Academic press, 2010.

[25] Ryohei Takahashi and Yuya Kajikawa. Computer-aided diagnosis: A survey with bibliometric analysis. *International Journal of Medical Informatics*, 101: 58–67, 2017.

[26] Nisreen I. Yassin, Shaimaa Omran, Enas M. El Houby, and Hemat Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: a systematic review. *Computer Methods and Programs in Biomedicine*, 2017.

[27] Bernd Jhne. *Digital Image processing: concepts, Algorithms, and Scientific Applications*. Springer, 1997.

[28] Laura Caponetti and Giovanna Castellano. *Fuzzy Logic for Image processing: A Gentle Introduction Using Java*. Springer, 2017.

[29] Walter Benjamin. A short history of photography. *Screen*, 13(1):5–26, 1972.

[30] Mark Harris. Snapping Up Kodak. *Spectrum, IEEE*, 51(2):30–62, 2014.

[31] Bernd Jahne, Horst Haussecker, and Peter Geissler. *Handbook of Computer Vision and Applications*, volume 2. Citeseer, 1999.

[32] John Semmlow and Benjamin Griffel. *Biosignal and Medical Image Processing*. CRC press, 2014.

[33] Wolfgang Birkfellner. *Applied Medical Image Processing: A Basic Course*. CRC Press, 2016.

[34] Kunio Doi. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31(4-5):198–211, 2007.

[35] Emilie L. Henriksen, Jonathan F. Carlsen, Ilse M. Vejborg, Michael B. Nielsen, and Carsten A. Lauridsen. The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review. *Acta Radiologica*, pages 13–18, 2018.

[36] Owen Saxton. *Computer Techniques for Image Processing in Electron Microscopy*, volume 10. Academic Press, 2013.

[37] Cornelis De Jager and Hans Nieuwenhuijzen. *Image processing techniques in astronomy: proceedings of a conference held in Utrecht on march 25–27, 1975*, volume 54. Springer Science & Business Media, 2012.

[38] Jian Gou Liu and Philippa Mason. *Essential Image Processing and GIS for Remote Sensing*. John Wiley & Sons, 2013.

[39] Gian Luca Foresti, Petri Mahonen, and Carlo Regazzoni. *Multimedia video-based surveillance systems: requirements, issues and solutions*, volume 573. Springer Science & Business Media, 2012.

[40] Graeme Jones, Nikos Paragios, and Carlo Regazzoni. *Video-based surveillance systems: computer vision and distributed processing*. Springer Science & Business Media, 2012.

[41] Hsu-Yung Cheng, Chih-Chia Weng, and Yi-Ying Chen. Vehicle detection in aerial surveillance using dynamic Bayesian networks. *IEEE Transactions on Image Processing*, 21(4):2152, 2012.

[42] Chi-hau Chen. *Signal and Image Processing for Remote Sensing*. CRC press, 2012.

[43] Gustavo Camps-Valls, Devis Tuia, Luis Gomez-Chova, Sandra Jimenez, and Jesus Malo. Remote sensing image processing. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 5(1):1–192, 2011.

[44] Thomas Lillesand, Ralph Kiefer, and Jonathan Chipman. *Remote sensing and image interpretation*. John Wiley & Sons, 2014.

[45] Daniel Guyer, Gaines Miles, Marvin Schreiber, Owen Mitchell, and Vern Vanderbilt. Machine vision and image processing for plant identification. *Transactions of the ASAE*, 29(6):1500–1507, 1986.

[46] James Cope, David Corney, Jonathan Clark, Paolo Remagnino, and Paul Wilkin. Plant species identification using digital morphometrics: a review. *Expert Systems with Applications*, 39(8):7562–7573, 2012.

[47] Tomas Lozano-Perez. *Autonomous Robot Vehicles*. Springer Science & Business Media, 2012.

[48] Jackie Courtney, Michael Magee, and Jagdishkumar Aggarwal. Robot guidance using computer vision. *Pattern Recognition*, 17(6):585–592, 1984.

[49] Johann Borenstein, Hobart Everett, Liquiang Feng, and David Wehe. Mobile robot positioning: sensors and techniques. *Journal of Robotic Systems*, 14 (4):231–249, 1997.

[50] Yucheng Fu and Yang Liu. Development of a robust image processing technique for bubbly flow measurement in a narrow rectangular channel. *International Journal of Multiphase Flow*, 84:217–228, 2016.

[51] Frans. Nieuwstadt. *Flow Visualization and Image Analysis*, volume 14. Springer Science & Business Media, 2012.

[52] Wolgang Merzkirch. *Flow Visualization*. Elsevier, 2012.

[53] Wilhem Burger and Mark Burge. *Digital Image Processing: An Algorithmic Introduction Using Java*. Springer, 2016.

[54] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, 1979.

[55] Pei-Gee Peter Ho. Image segmentation by autoregressive time series model. In *Image Segmentation*. InTech, 2011.

[56] Shubha Sanu and Pushpa Tamase. Satellite image mining using content based image retrieval. *International Journal of Engineering Science and Computing*, 7(7):13928–13931, 2017.

[57] Weihong Cui, Zequn Guan, and Zhiyi Zhang. An improved region growing algorithm for image segmentation. In *2008 International Conference on Computer Science and Software Engineering*, volume 6, pages 93–96. IEEE, 2008.

[58] Yuri Boykov and Gareth Funka-Lea. Graph cuts and efficient ND image segmentation. *International Journal of Computer Vision*, 70(2):109–131, 2006.

[59] Yuri Boykov and Marie-Pierre Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *Proceedings*

Eighth IEEE International Conference on Computer Vision, ICCV 2001, volume 1, pages 105–112. IEEE, 2001.

[60] Ahmad Ali Abin, Mehran Fotouhi, and Shohreh Kasaei. Skin segmentation based on cellular learning automata. In *Proceedings of the 6th International Conference on Advances in Mobile Computing and Multimedia*, pages 254–259. ACM, 2008.

[61] Robert Schalkoff. *Digital Image Processing and Computer Vision*, volume 286. Wiley New York, 1989.

[62] Chen Chi Hau. *Handbook of Pattern Recognition and Computer Vision*. World Scientific, 2015.

[63] Wenan Chen, Rebecca Smith, Soo-Yeon Ji, Kevin Ward, and Kayvan Najarian. Automated ventricular systems segmentation in brain CT images by combining low-level segmentation and high-level template matching. *BMC Medical Informatics and Decision Making*, 9(1):1–14, 2009.

[64] Luc Devroye, Laszlo Gyorfi, and Gabor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

[65] John Cowell. Syntactic pattern recognizer for vehicle identification numbers. *Image and Vision Computing*, 13(1):13–19, 1995.

[66] Luis Gomez-Chova, Devis Tuia, Gabriele Moser, and Gustau Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, 2015.

[67] Christian Koch, Kristina Georgieva, Varun Kasireddy, Burcu Akinci, and Paul Fieguth. A review on computer vision based defect detection and con-

dition assessment of concrete and asphalt civil infrastructure. *Advanced Engineering Informatics*, 29(2):196–210, 2015.

[68] Biswa Ranjan Acharya and Pradosh Kumar Gantayat. Recognition of human unusual activity in surveillance videos. *International Journal of Research and Scientific Innovation IJRSI*, 2:18–23, 2015.

[69] Carlos Leon, Lyudmila Mihaylova, and Hans Driessen. Tracking of interacting targets. In *2017 20th International Conference on Information Fusion*, pages 1–8. IEEE, 2017.

[70] Amin Raid, Wael Khedr, Ma El-Dosuky, and Wesam Ahmed. Jpeg image compression using discrete cosine transform- A survey. *arXiv preprint arXiv:1405.6147*, 2014.

[71] Peng Cao, Xiaoli Liu, Jinzhu Yang, Dazhe Zhao, Wei Li, Min Huang, and Osmar Zaiane. A multi-kernel based framework for heterogeneous feature selection and over-sampling for computer-aided detection of pulmonary nodules. *Pattern Recognition*, 64:327–346, 2017.

[72] Peter Huang, Seyoun Park, Rongkai Yan, Junghoon Lee, Linda Chu, Cheng Lin, Amira Hussien, Joshua Rathmell, Brett Thomas, and Chen Chen. Added value of computer-aided CT image features for early lung cancer diagnosis with small pulmonary nodules: a matched case-control study. *Radiology*, 286(1):286–295, 2017.

[73] Sebastian Mika, Gunnar Ratsch, Jason Weston, Bernhard Scholkopf, and Klaus-Robert Mullers. Fisher discriminant analysis with kernels. In *Proceedings of the 1999 IEEE Signal Processing Society Workshop, Neural Networks for Signal Processing IX*, pages 41–48. IEEE, 1999.

[74] Guohui Zhou, Yuanyuan Wang, and Weiqi Wang. Diagnosis of hepatic fibrosis by ultrasonic image analysis. In *2012 International Conference on Biomedical Engineering and Biotechnology ICBEB*, pages 775–776. IEEE, 2012.

[75] Bernhard. Boser, Isabelle. Guyon, and Vladimir. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152. ACM, 1992.

[76] Surendra Kumar Rakse and Sanyam Shukla. Spam classification using new kernel function in support vector machine. *International Journal on Computer Science and Engineering*, 2(5):2010, 1819.

[77] Ruey-Feng Chang, Wen-Jie Wu, Woo Kyung Moon, and Dar-Ren Chen. Automatic ultrasound segmentation and morphology based diagnosis of solid breast tumors. *Breast Cancer Research and Treatment*, 89(2):179, 2005.

[78] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[79] Nitish Kumar, Pinki Kumari, Preetish Ranjan, and Abhishek Vaish. ARIMA model based breast cancer detection and classification through image processing. In *2014 Students Conference on Engineering and Systems SCES*, pages 1–5. IEEE, 2014.

[80] Nasir Ahmed, T. Natarajan, and K. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 100(1):90–93, 1974.

[81] Florian Wamser, Pedro Casas, Michael Seufert, Christian Moldovan, Phuoc Tran-Gia, and Tobias Hossfeld. Modeling the YouTube stack: From packets to quality of experience. *Computer Networks*, 109:211–224, 2016.

[82] John Watkinson. *The MPEG Handbook*. Focal Press, 2012.

[83] Meng Joo Er, Weilong Chen, and Shiqian Wu. High-speed face recognition based on discrete cosine transform and RBF neural networks. *IEEE Transactions on Neural Networks*, 16(3):679–691, 2005.

[84] Samuel Lukas, Aditya Rama Mitra, Ririn Ikana Desanti, and Dion Krisnadi. Student attendance system in classroom using face recognition technique. In *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pages 1032–1035. IEEE, 2016.

[85] Ziad M. Hafed and Martin D. Levine. Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, 43(3):167–188, 2001.

[86] Bo Shen and Ishwar K. Sethi. Direct feature extraction from compressed images. In *Storage and Retrieval for Still Image and Video Databases IV*, volume 2670, pages 404–415. International Society for Optics and Photonics, 1996.

[87] Mehran Kafai, Kave Eshghi, and Bir Bhanu. Discrete cosine transform locality-sensitive hashes for face retrieval. *IEEE Transactions on Multimedia*, 16(4):1090–1103, 2014.

[88] Stephen A. Billings. *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-temporal Domains*. John Wiley & Sons, 2013.

[89] Sreejit Chakravarty and Pradipta K. Dash. A PSO based integrated functional link net and interval type-2 fuzzy logic system for predicting stock market indices. *Applied Soft Computing*, 12(2):931–941, 2012.

[90] Ihsan Yassin, Muhammad Abdul Khalid, Sukreen Herman, Ibrahim Pasya, Norfishah Wahab, and Zaiki Awang. Multi-layer perceptron (MLP)-based nonlinear auto-regressive with exogenous inputs (NARX) stock forecasting model. *International Journal on Advanced Science, Engineering and Information Technology*, 7(3):1098–1103, 2017.

[91] Jayashree Chadalawada, Vojtech Havlicek, and Vladan Babovic. A genetic programming approach to system identification of rainfall-runoff models. *Water Resources Management*, 31(12):3975–3992, 2017.

[92] Rui Huang, Tiana Huang, Rajit Gadh, and Na Li. Solar generation prediction using the ARMA model in a laboratory-level micro-grid. In *2012 IEEE Third International Conference on Smart Grid Communications SmartGridComm*, pages 528–533. IEEE, 2012.

[93] Kah Wai Cheah and Noor Atinah Ahmad. Fuzzy recursive least-squares approach in speech system identification:a transformed domain LPC model. *International Journal of Electrical and Computer Engineering IJECE*, 7(2): 842–849, 2017.

[94] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP*, pages 6645–6649. IEEE, 2013.

[95] Mohammed Falah Mohammed and Chee Peng Lim. An enhanced fuzzy min–max neural network for pattern classification. *IEEE Transactions on Neural Networks and Learning Systems*, 26(3):417–429, 2015.

[96] Sankar K. Pal and Paul P. Wang. *Genetic Algorithms for Pattern Recognition*. CRC press, 2017.

[97] Eugene A. Morelli and Vladislav Klein. *Aircraft System Identification: Theory and Practice*. Sunflyte Enterprises Williamsburg, VA, 2016.

[98] Joshua Harris, Frank Arthurs, James V. Henrickson, and John Valasek. Aircraft system identification using artificial neural networks with flight test data. In *2016 International Conference on Unmanned Aircraft Systems ICUAS*, pages 679–688. IEEE, 2016.

[99] Lennart Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1–12, 2010.

[100] Stephen Billings. Identification of nonlinear systems using parameter estimation techniques. In *Institute of Electrical Engineers Conference*, pages 183–190, 1981.

[101] Quanmin Zhu and Stephen Billings. Parameter estimation for stochastic nonlinear rational models. *International Journal of Control*, 57(2):309–333, 1993.

[102] Martin Schetzen. The Volterra and Wiener theories of nonlinear systems. *Proceedings of the IEEE*, 69(12):1557–1573, 1980.

[103] Johan Schoukens, Liesbeth Gomme, Wendy Van Moer, and Yves Rolain. Identification of a block-structured nonlinear feedback system, applied to

a microwave crystal detector. *IEEE Transactions on Instrumentation and Measurement*, 57(8):1734–1740, 2008.

[104] David Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

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

[106] Charles K. Chui. *An introduction to Wavelets*. Elsevier, 2016.

[107] Didier Dubois and Henri Prade. *Fundamentals of Fuzzy Sets*, volume 7. Springer Science & Business Media, 2012.

[108] Janaina Schwarzrock, Iulisloi Zacarias, Ana Bazzan, Ricardo Queiroz de Araujo Fernandes, Leonardo Henrique Moreira, and Edison Pignaton de Freitas. Solving task allocation problem in multi unmanned aerial vehicles systems using swarm intelligence. *Engineering Applications of Artificial Intelligence*, 72:10–20, 2018.

[109] Tobias Munker and Oliver Nelles. Nonlinear system identification with regularized local FIR model networks. *IFAC-PapersOnLine*, 49(5):61–66, 2016.

[110] Ion Leontaritis and Stephen Billings. Input-output parametric models for non-linear systems part I: deterministic non-linear systems. *International Journal of Control*, 41(2):303–328, 1985.

[111] Ronald K. Pearson. *Discrete-Time Dynamic Models*. Oxford University Press, 1999.

[112] Hua-Liang Wei, Stephen Billings, Yifan Zhao, and Lingzhong Guo. Lattice dynamical wavelet neural networks implemented using particle swarm optimization for spatio–temporal system identification. *IEEE Transactions on Neural Networks*, 20(1):181–185, 2009.

[113] Gonzalo Acuna, Cristian Ramirez, and Millaray Curilem. Comparing NARX and NARMAX models using ANN and SVM for cash demand forecasting for ATM. In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2012.

[114] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1994.

[115] Paul J. Werbos. Applications of advances in nonlinear sensitivity analysis. In *System Modeling and Optimization*, pages 762–770. Springer, 1982.

[116] John Moody and Christian Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[117] Ian Witten, Eibe Frank, Mark Hall, and Christopher Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[118] Anil Jain, Robert Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

[119] Anil Jain, Jianchang Mao, and K. Mohiuddin. Artificial neural networks: A tutorial. *Computer*, 29(3):31–44, 1996.

[120] Zhexue Huang. Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD*, pages 21–34. Singapore, 1997.

[121] Rajat Raina, Anand Madhavan, and Andrew Ng. Large-scale deep unsupervised learning using graphics processors. In *Proceedings of the 26th Annual InternationalConference on Machine Learning*, pages 873–880. ACM, 2009.

[122] Scott Martin. How a deep neural network sees, September 2018. URL `https://blogs.nvidia.com/blog/2018/09/05/whats-the-difference-between-a-cnn-and-an-rnn/`. [Online; accessed October 20, 2018].

[123] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[124] Carlos Affonso, Andre Rossi, Fabio Vieira, and Andre Leon-Ferreira. Deep learning for biological image classification. *Expert Systems with Applications*, 85:114–122, 2017.

[125] Hamid Beigy and Mohamad Reza Meybodi. A learning automata-based algorithm for determination of the number of hidden units for three-layer neural networks. *International Journal of Systems Science*, 40(1):101–118, 2009.

[126] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Fredo Durand, and William T. Freeman. The visual microphone: passive recovery of sound from video. *ACM Transactions on Graphics TOG*, 33(4):79, 2014.

[127] Raia Hadsell, Pierre Sermanet, Jan Ben, A. Erkan, Jeff Han, Beat Flepp, Urs Muller, and Yann LeCun. Online learning for offroad robots: Using spatial label propagation to learn long-range traversability. In *Proceedings of Robotics: Science and Systems (RSS)*, volume 11, page 32, 2007.

[128] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. *arXiv preprint arXiv:1302.4389*, 2013.

[129] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[130] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE Transactions Audio, Speech & Language Processing*, 20(1):14–22, 2012.

[131] George E. Dahl, Dong Yu, Li Deng, and Alex Acero. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, 2012.

[132] Liang Lu, Xingxing Zhang, Kyunghyun Cho, and Steve Renals. A study of the recurrent neural network encoder-decoder for large vocabulary speech recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*. International Speech and Communication Association, Citeseer, 2015.

[133] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[134] Tomavs Kovcisky, Karl Moritz Hermann, and Phil Blunsom. Learning bilingual word representations by marginalizing alignments. *arXiv preprint arXiv:1405.0947*, 2014.

[135] Ronan Collobert. Deep learning for efficient discriminative parsing. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 224–232, 2011.

[136] Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(8):2493–2537, 2011.

[137] Yann LeCun. Generalization and network design strategies. *Connectionism in Perspective*, 19:143–155, 1989.

[138] Charles Cadieu, Ha Hong, Daniel Yamins, Nicolas Pinto, Diego Ardila, Ethan Solomon, Najib Majaj, and James DiCarlo. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLOS Computational Biology*, 10(12):e1003963, 2014.

[139] Yi-Tong Zhou and Rama Chellappa. Computation of optical flow using a neural network. In *IEEE International Conference on Neural Networks*, volume 1998, pages 71–78, 1988.

[140] Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014.

[141] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne E. Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990.

[142] Alexander Waibel, Toshiyuki Hanazawa, Geoffrey Hinton, Kiyohiro Shikano, and Kevin Lang. Phoneme recognition using time-delay neural networks. In *Readings in Speech Recognition*, pages 393–404. Elsevier, 1990.

[143] Patrice Simard, Dave Steinkraus, and John Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Seventh International Conference on Document Analysis and Recognition, Proceedings*, volume 3, pages 958–963. IEEE, Citeseer, 2003.

[144] Chenglong Wang, Feijun Jiang, and Hongxia Yang. A hybrid framework for text modeling with convolutional RNN. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2061–2069. ACM, 2017.

[145] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1891–1898, 2014.

[146] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4305–4314, 2015.

[147] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[148] Tiago Matias, Francisco Souza, Rui Araújo, and Carlos Henggeler Antunes. Learning of a single-hidden layer feedforward neural network using an optimized extreme learning machine. *Neurocomputing*, 129:428–436, 2014.

[149] Guang-Bin Huang and Chee Kheong Siew. Extreme learning machine: RBF network case. In *ICARCV*, volume 2, pages 1029–1036, 2004.

[150] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[151] Guang-Bin Huang and Chee-Kheong Siew. Extreme learning machine with randomly assigned RBF kernels. *International Journal of Information Technology*, 11(1):16–24, 2005.

[152] Yoshua Bengio et al. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[153] Kai Sun, Jiangshe Zhang, Chunxia Zhang, and Junying Hu. Generalized extreme learning machine autoencoder and a new deep neural network. *Neurocomputing*, 230:374–381, 2017.

[154] Liyanaarachchi Lekamalage Chamara Kasun, Hongming Zhou, Guang-Bin Huang, and Chi Man Vong. Representational learning with extreme learning machine for big data. *IEEE Intelligent Systems*, 28(6):31–34, 2013.

[155] George Cybenko. Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2:183–192, 1989.

[156] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.

[157] Jooyoung Park and Irwin W. Sandberg. Approximation and radial-basis-function networks. *Neural Computation*, 5(2):305–316, 1993.

[158] Sheng Chen and Stephen Billings. Neural networks for nonlinear dynamic system modelling and identification. *International Journal of Control*, 56(2): 319–346, 1992.

[159] Meng Joo Er, Shiqian Wu, Juwei Lu, and Hock Lye Toh. Face recognition with radial basis function (RBF) neural networks. *IEEE Transactions on Neural Networks*, 13(3):697–710, 2002.

[160] Tomaso Poggio and Shimon Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343(6255):263, 1990.

[161] Rui Yang, Poi Voon Er, Zidong Wang, and Kok Kiong Tan. An RBF neural network approach towards precision motion system with selective sensor fusion. *Neurocomputing*, 199:31–39, 2016.

[162] Mellisa Pratiwi, Jeklin Harefa, Sakka Nanda, et al. Mammograms classification using gray-level co-occurrence matrix and radial basis function neural network. *Procedia Computer Science*, 59:83–91, 2015.

[163] Zhihai Lu, Siyuan Lu, Ge Liu, Yudong Zhang, Jianfei Yang, and Preetha Phillips. A pathological brain detection system based on radial basis function neural network. *Journal of Medical Imaging and Health Informatics*, 6(5): 1218–1222, 2016.

[164] Vinay Kumar, Abul Abbas, Nelson Fausto, and Jon Aster. *Robbins and Cotran Pathologic Basis of Disease, Professional Edition*. Elsevier Health Sciences, 2014.

[165] Jerry M. Mendel. *Uncertain Rule-Based Fuzzy Systems*. Springer, 2017.

[166] Krassimir T. Atanassov. Type-1 fuzzy sets and intuitionistic fuzzy sets. *Algorithms*, 10(3):106, 2017.

[167] Thanh Nguyen, Abbas Khosravi, Douglas Creighton, and Saeid Nahavandi. Classification of healthcare data using genetic fuzzy logic system and wavelets. *Expert Systems with Applications*, 42(4):2184–2197, 2015.

[168] Balu Bhasuran, Gurusamy Murugesan, Sabenabanu Abdulkadhar, and Jeyakumar Natarajan. Stacked ensemble combined with fuzzy matching for biomedical named entity recognition of diseases. *Journal of Biomedical Informatics*, 64:1–9, 2016.

[169] Guichao Fan, Denghua Zhong, Fugen Yan, and Pan Yue. A hybrid fuzzy evaluation method for curtain grouting efficiency assessment based on an AHP method extended by D numbers. *Expert Systems with Applications*, 44: 289–303, 2016.

[170] Ning Wang, Meng Joo Er, Jing-Chao Sun, and Yan-Cheng Liu. Adaptive robust online constructive fuzzy control of a complex surface vehicle system. *IEEE Transactions on Cybernetics*, 46(7):1511–1523, 2016.

[171] Adrian Rubio Solis and George Panoutsos. Granular computing neural-fuzzy modelling: A neutrosophic approach. *Applied Soft Computing*, 13(9):4010–4021, 2013.

[172] Luis Aguirre and Stephen Billings. Improved structure selection for nonlinear models based on term clustering. *International Journal of Control*, 62(3): 569–587, 1995.

[173] Samir Angelo Martins, Erivelton Nepomuceno, and Marcio Falcao Santos Barroso. Improved structure detection for polynomial NARX models using a multiobjective error reduction ratio. *Journal of Control, Automation and Electrical Systems*, 24(6):764–772, 2013.

[174] Mohd Z. Zakaria, Hishamuddin Jamaluddin, Robiah Ahmad, and Sayed M. R. Loghmanian. Comparison between multi-objective and single-objective optimization for the modeling of dynamic systems. *Proceedings of the In-*

stitution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 226(7):994–1005, 2012.

[175] Hua-Liang Wei, Stephen Billings, and Jianhua Liu. Term and variable selection for non-linear system identification. *International Journal of Control*, 77(1):86–110, 2004.

[176] Yuzhu Guo, Ling Zhong Guo, Stephen A. Billings, and Hua-Liang Wei. Identification of nonlinear systems with non-persistent excitation using an iterative forward orthogonal least squares regression algorithm. *International Journal of Modelling, Identification and Control*, 23(1):1–7, 2015.

[177] Kezhi Mao and Stephen Billings. Algorithms for minimal model structure detection in nonlinear dynamic system identification. *International Journal of Control*, 68(2):311–330, 1997.

[178] Jinyao Yan and John Deller. NARMAX model identification using a set-theoretic evolutionary approach. *Signal Processing*, 123:30–41, 2016.

[179] Oliver Nelles, Alexander Fink, and Rolf Isermann. Local linear model trees (LOLIMOT) toolbox for nonlinear system identification. *IFAC Proceedings Volumes*, 33(15):845–850, 2000.

[180] Sunil L. Kukreja, Johan Lofberg, and Martin J. Brenner. A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proceedings Volumes*, 39(1):814–819, 2006.

[181] Brian Karrer and Mark E. J. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[182] Jan Puzicha, Thomas Hofmann, and Joachim M. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 33(4):617–634, 2000.

[183] Stephen Billings, Michael Korenberg, and Sheng Chen. Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *International Journal of Systems Science*, 19(8):1559–1568, 1988.

[184] Sheng Chen, Stephen A. Billings, and Wan Luo. Orthogonal least squares methods and their application to non-linear system identification. *International Journal of Control*, 50(5):1873–1896, 1989.

[185] Michael Korenberg, Stephen Billings, Jeh-Ping Liu, and Patrick McIlroy. Orthogonal parameter estimation algorithm for non-linear stochastic systems. *International Journal of Control*, 48(1):193–210, 1988.

[186] John G. F. Francis. The QR transformation a unitary analogue to the LR transformation - Part 1. *The Computer Journal*, 4(3):265–271, 1961.

[187] Steven J. Leon, AAke Bjrck, and Walter Gander. Gram-Schmidt orthogonalization: 100 years and more. *Numerical Linear Algebra with Applications*, 20(3):492–532, 2013.

[188] Stephen Billings, Michael Korenberg, and Sheng Chen. Identification of MIMO non-linear systems using a forward-regression orthogonal estimator. *International journal of control*, 49(6):2157–2189, 1989.

[189] Sheng Chen, Colin F. N. Cowan, and Peter M. Grant. Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, 2(2):302–309, 1991.

[190] Danlin Yu, Barry Gomm, and David Williams. A recursive orthogonal least squares algorithm for training RBF networks. *Neural Processing Letters*, 5 (3):167–176, 1997.

[191] Li-Xin. Wang and Jerry M. Mendel. Fuzzy basis functions, universal approximation, and orthogonal least-squares learning. *IEEE transactions on Neural Networks*, 3(5):807–814, 1992.

[192] Xia. Hong, Martin. Brown, Sheng. Chen, and Christopher Harris. Sparse model identification using orthogonal forward regression with basis pursuit and D-optimality. *IEE Proceedings-Control Theory and Applications*, 151(4): 491–498, 2004.

[193] Kunio Doi. Current status and future potential of computer-aided diagnosis in medical imaging. *The British Journal of Radiology*, 78(suppl_1):s3–s19, 2005.

[194] Irene Pollanen, Billy Braithwaite, Tiia Ikonen, Harri Niska, Keijo Haataja, Pekka Toivanen, and Teemu Tolonen. Computer-aided breast cancer histopathological diagnosis: Comparative analysis of three DTOCS-based features: SW-DTOCS, SW-WDTOCS and SW-3-4-DTOCS. In *2014 4th International Conference on Image Processing Theory, Tools and Applications IPTA*, pages 1–6. IEEE, 2014.

[195] Ahmedin Jemal, Freddie Bray, Melissa Center, Jacques Ferlay, Elizabeth Ward, and David Forman. Global cancer statistics. *CA: a Cancer Journal for Clinicians*, 61(2):69–90, 2011.

[196] Jinshan Tang, Rangaraj M. Rangayyan, Jun Xu, Issam El Naqa, and Yongyi Yang. Computer-aided detection and diagnosis of breast cancer with mam-

mography: recent advances. *IEEE Transactions on Information Technology in Biomedicine*, 13(2):236–251, 2009.

[197] Nisreen I. Yassin, Shaimaa Omran, Enas M. El Houby, and Hemat Allam. Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. *Computer Methods and Programs in Biomedicine*, 156:25–45, 2018.

[198] David Martin Powers. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[199] Etienne von Lavante and J. Alison Noble. Segmentation of breast cancer masses in ultrasound using radio-frequency signal derived parameters and strain estimates. In *2008. ISBI 2008. 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 536–539. IEEE, 2008.

[200] Burak Alacam, Birsen Yazici, and Nihat M. Bilgutay. Breast tissue characterization based on fractional differencing model of ultrasonic RF echo. In *Medical Imaging 2003: Ultrasonic Imaging and Signal Processing*, volume 5035, pages 460–471. International Society for Optics and Photonics, 2003.

[201] Liyang Wei, Yongyi Yang, Robert M. Nishikawa, and Yulei Jiang. A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications. *IEEE Transactions on Medical Imaging*, 24(3): 371–380, 2005.

[202] Burak Alacam, Birsen Yazici, Nihat Bilgutay, Flemming Forsberg, and Catherine Piccoli. Breast tissue characterization using FARMA modeling of ultrasonic RF echo. *Ultrasound in Medicine & Biology*, 30(10):1397–1407, 2004.

[203] Nidhal Bouaynaya, Jerzy Zielinski, and Dan Schonfeld. Two-dimensional ARMA modeling for breast cancer detection and classification. *arXiv preprint arXiv:0906.3722*, 2009.

[204] Satya P. Singh and Shabana Urooj. An improved CAD system for breast cancer diagnosis based on generalized pseudo-Zernike moment and Ada-DEWNN classifier. *Journal of Medical Systems*, 40(4):105, 2016.

[205] Weiying Xie, Yunsong Li, and Yide Ma. Breast mass classification in digital mammography based on extreme learning machine. *Neurocomputing*, 173: 930–941, 2016.

[206] Saeid Asgari Taghanaki, Jeremy Kawahara, Brandon Miles, and Ghassan Hamarneh. Pareto-optimal multi-objective dimensionality reduction deep auto-encoder for mammography classification. *Computer Methods and Programs in Biomedicine*, 145:85–93, 2017.

[207] Anant Madabhushi and George Lee. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170–175, 2016.

[208] Hao Wu and Saurabh Prasad. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259–1270, 2018.

[209] Fan Liu, Jinhui Tang, Yan Song, Ye Bi, and Sai Yang. Local structure based multi-phase collaborative representation for face recognition with single sample per person. *Information Sciences*, 346:198–215, 2016.

[210] Quentin F. Stout. Supporting divide-and-conquer algorithms for image processing. *Journal of Parallel and Distributed Computing*, 4(1):95–115, 1987.

[211] Shane O'Grady and Maria Morgan. Microcalcifications in breast cancer: From pathophysiology to diagnosis and prognosis. *Biochimica et Biophysica Acta BBA-Reviews on Cancer*, 2018.

[212] John Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, and S. Kok. The mammographic image analysis society digital mammogram database. In *Exerpta Medica. International Congress Series*, volume 1069, pages 375–378, 1994.

[213] Thomas Pröll and M. Nazmul Karim. Model-predictive pH control using real-time NARX approach. *AIChE Journal*, 40(2):269–282, 1994.

[214] Azlee Zabidi, Nooritawati Md Tahir, Ihsan Mohd Yassin, and Zairi Ismael Rizman. The performance of binary artificial bee colony (BABC) in structure selection of polynomial NARX and NARMAX models. *International Journal on Advanced Science, Engineering and Information Technology*, 7(2):373–379, 2017.

[215] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[216] Mario Mustra and Mislav Grgic. Robust automatic breast and pectoral muscle segmentation from scanned mammograms. *Signal Processing*, 93(10): 2817–2827, 2013.

[217] Carlos Beltran Perez and Hua-Liang Wei. Digital image classification and detection using a 2D-NARX model. In *2017 23rd International Conference on Automation and Computing ICAC*, pages 1–6. IEEE, 2017.

[218] Christopher J. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[219] William W. Hsieh. *Machine learning methods in the environmental sciences: Neural networks and kernels*. Cambridge university press, 2009.

[220] Bernhard Schölkopf, Alexander J. Smola, and Francis Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[221] P. Schiilkop, Chris Burgest, and Vladimir Vapnik. Extracting support data for a given task. In *Proceedings of the 1st international conference on knowledge discovery & data mining*, pages 252–257, 1995.

[222] Stephen A. Billings, Hua-Liang Wei, and Michael A. Balikhin. Generalized multiscale radial basis function networks. *Neural Networks*, 20(10):1081–1094, 2007.

[223] Tomaso Poggio and Federico Girosi. A theory of networks for approximation and learning. Technical report, Massachusetts Institute of Technology Cambridge Artificial Intelligence Laboratory, 1989.

[224] Wojtek Krzanowski and Yen-Ting Lai. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, 44 (1):23–34, 1988.

[225] Stephen Billings, Michael Korenberg, and Sheng Chen. Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *International Journal of Systems Science*, 19(8):1559–1568, 1988.

[226] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[227] Raouf Boutaba, Mohammad A. Salahuddin, Noura Limam, Sara Ayoubi, Nashid Shahriar, Felipe Estrada-Solano, and Oscar M. Caicedo. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities. *Journal of Internet Services and Applications*, 9 (1):16, 2018.

[228] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, pages 2035–2043, 2009.

[229] C. Beltran Perez and H. Wei. Image Classification Using Generalized Multiscale RBF Networks and Discrete Cosine Transform. In *Proceedings of the 24th International Conference on Automation and Computing*. IEEE, 2018.

[230] Charles E. Metz. ROC methodology in radiologic imaging. *Investigative radiology*, 21(9):720–733, 1986.

[231] Richard Wender, Elizabeth T. Fontham, Ermilo Barrera Jr, Graham A. Colditz, Timothy R. Church, David S. Ettinger, Ruth Etzioni, Christopher R. Flowers, G. Scott Gazelle, Douglas K. Kelsey, et al. American Cancer Society lung cancer screening guidelines. *CA: A Cancer Journal for Clinicians*, 63 (2):106–117, 2013.

[232] Jun Xu, Lei Xiang, Qingshan Liu, Hannah Gilmore, Jianzhong Wu, Jinghai Tang, and Anant Madabhushi. Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. *IEEE Transactions on Medical Imaging*, 35(1):119–130, 2016.

[233] Wu Jun, Wang Shitong, and Fu-lai Chung. Positive and negative fuzzy rule system, extreme learning machine and image classification. *International Journal of Machine Learning and Cybernetics*, 2(4):261–271, 2011.

[234] Rahimeh Rouhi, Mehdi Jafari, Shohreh Kasaei, and Peiman Keshavarzian. Benign and malignant breast tumors classification based on region growing and CNN segmentation. *Expert Systems with Applications*, 42(3):990–1002, 2015.

[235] John Arevalo, Fabio A. González, Raúl Ramos-Pollán, Jose L. Oliveira, and Miguel Angel Guevara Lopez. Representation learning for mammography mass lesion classification with convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 127:248–257, 2016.

[236] Petra Vidnerová and Roman Neruda. Deep Networks with RBF Layers to Prevent Adversarial Examples. In *International Conference on Artificial Intelligence and Soft Computing*, pages 257–266. Springer, 2018.

[237] Guang-Bin Huang, Qin-Yu Zhu, and Chee-Kheong Siew. Extreme learning machine: theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006.

[238] Guang-Bin Huang and Chee Kheong Siew. Extreme learning machine: RBF network case. In *ICARCV*, volume 2, pages 1029–1036, 2004.

[239] J.-S. Jang and C.-T. Sun. Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural Networks*, 4(1):156–159, 1993.

[240] Adrian Rubio-Solis and George Panoutsos. Interval type-2 radial basis function neural network: A modeling framework. *IEEE Transactions on Fuzzy Systems*, 23(2):457–473, 2015.

[241] Adrian Rubio-Solis, Patricia Melin, Uriel Martinez-Hernandez, and George Panoutsos. General Type-2 Radial Basis Function Neural Network: A Data-Driven Fuzzy Model. *IEEE Transactions on Fuzzy Systems*, 2018.

[242] Ali Baraka, George Panoutsos, and Stephen Cater. Perpetual Learning Framework based on Type-2 Fuzzy Logic System for a Complex Manufacturing Process. *IFAC-PapersOnLine*, 49(20):143–148, 2016.

[243] Uriel Martinez-Hernandez, Adrian Rubio-Solis, and Abbas A. Dehghani-Sanij. Recognition of walking activity and prediction of gait periods with a CNN and first-order MC strategy. In *2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics Biorob*, pages 897–902. IEEE, 2018.

[244] Jerry M. Mendel. General type-2 fuzzy logic systems made simple: a tutorial. *IEEE Transactions on Fuzzy Systems*, 22(5):1162–1182, 2014.

[245] A. Rubio-Solis, U. Martinez-Hernandez, and G. Panoutsos. Evolutionary extreme learning machine for the interval type-2 radial basis function neural network: a fuzzy modelling approach. In *World Congress on Computational Intelligence*. IEEE, 2018.

[246] Dongrui Wu and Jerry M. Mendel. Enhanced karnik–mendel algorithms. *IEEE Transactions on Fuzzy Systems*, 17(4):923–934, 2009.

[247] Jerry M. Mendel. Computing derivatives in interval type-2 fuzzy logic systems. *IEEE Transactions on Fuzzy Systems*, 12(1):84–98, 2004.

[248] Dorra Mellouli, Tarek M. Hamdani, and Adel M. Alimi. Deep neural network with RBF and sparse auto-encoders for numeral recognition. In *2015*

15th International Conference on Intelligent Systems Design and Applications ISDA, pages 468–472. IEEE, 2015.

[249] Chia-Feng Juang, Ren-Bo Huang, and Wei-Yuan Cheng. An interval type-2 fuzzy-neural network with support-vector regression for noisy regression problems. *IEEE Transactions on Fuzzy Systems*, 18(4):686–699, 2010.

[250] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[251] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[252] Phil Kim. *MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence*. Apress, 2017.

Appendix A

Examples of study case mammograms

Section 3.3.1 showed examples of 3 standard mammograms from X-ray scans of the mini-MIAS database [212], and in 3.3.4 the resulting models were reported for the same images. This annexe shows the images referred into a larger magnitude to facilitate their visualisation.

Please note that the images were converted into negative to reducing the ink waste in printing.

A.1 Benign mammogram mdb005

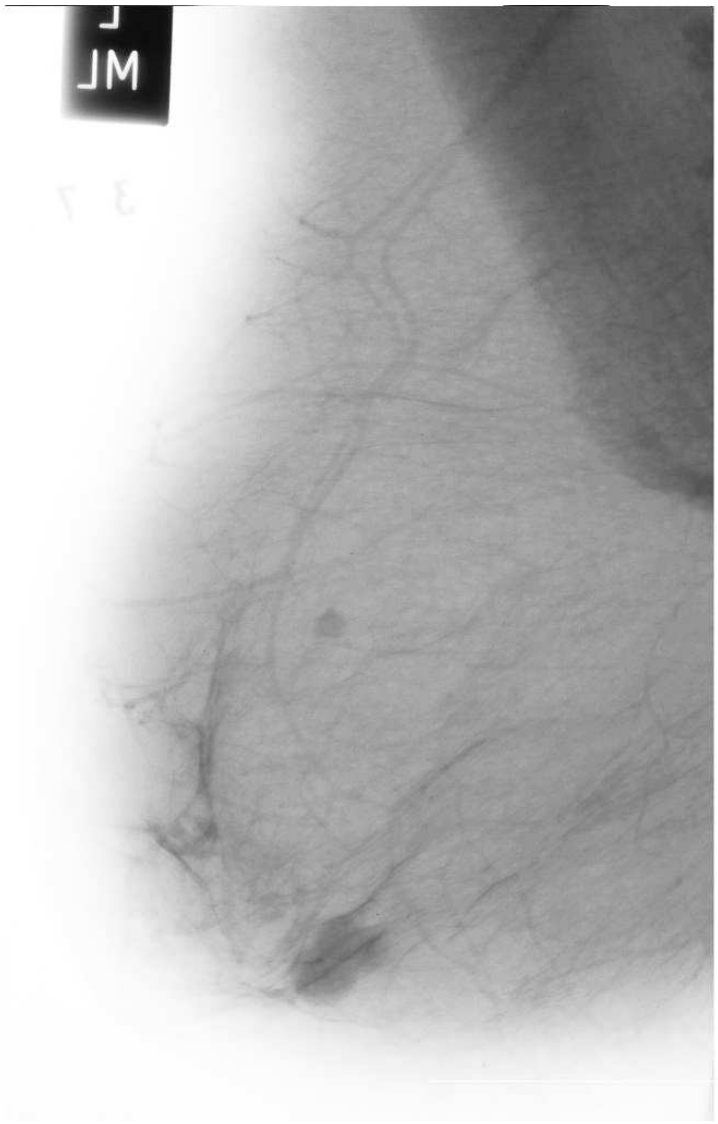


Figure A.1: Mammogram with a benign tumour. Film mdb005 from the mini-MIAS database [212].

A.2 Normal mammogram mdb009

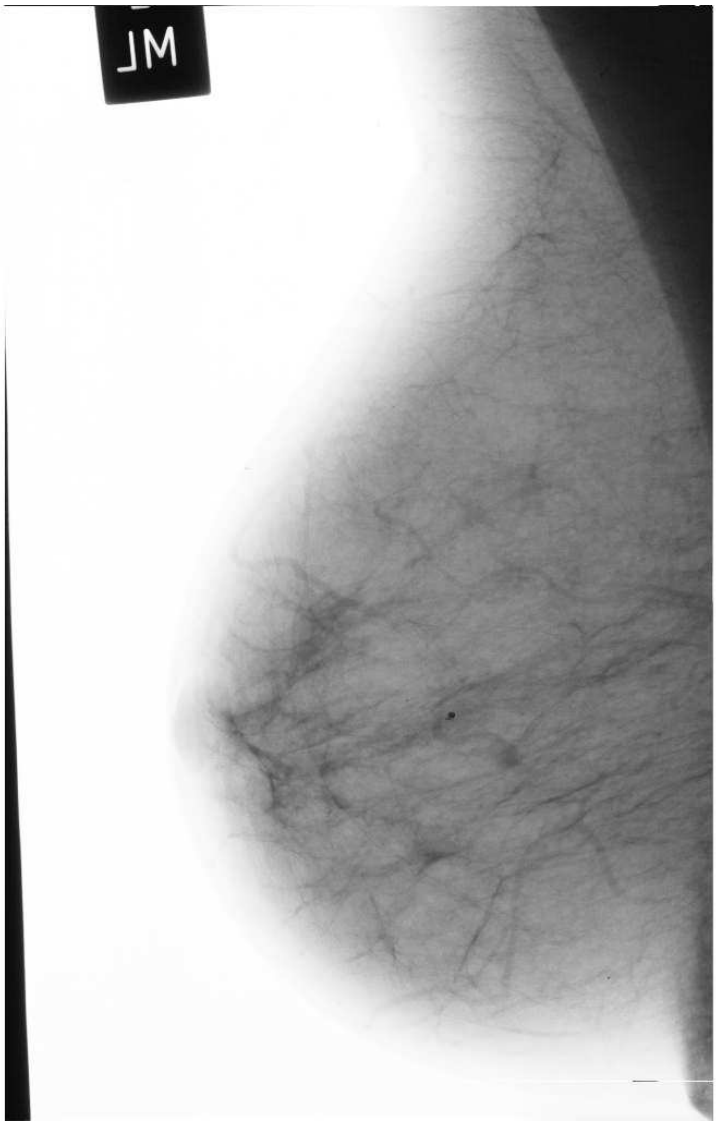


Figure A.2: Mammogram in healthy clinical condition. Film mdb009 from the mini-MIAS database [212].

A.3 Malign mammogram mdb028

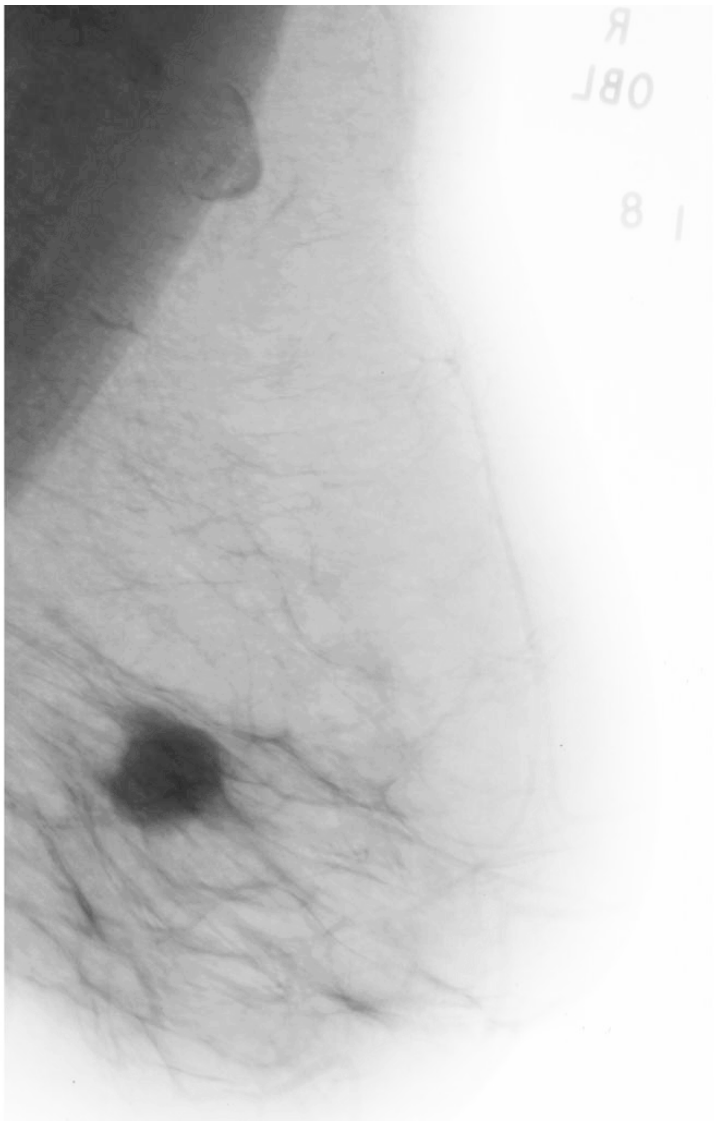


Figure A.3: Mammogram with a malign tumour. Film mdb028 from the mini-MIAS database [212].

Appendix B

2D NARX testing results

Table B.1: 2D NARX model test results 1-15

Test no.	Mamm. no.	Real condition	Result	Pred. type.
1	27	Normal	True negative	
2	75	Malign	True positive	
3	59	Benign	True positive	Benign
4	60	Normal	True negative	
5	25	Benign	True positive	Benign
6	28	Malign	True positive	Malign
7	76	Normal	True negative	
8	77	Normal	True negative	
9	78	Normal	True negative	
10	79	Normal	True negative	
11	80	Benign	True positive	Benign
12	83	Benign	True positive	Benign
13	84	Normal	True negative	
14	87	Normal	False positive	
15	88	Normal	True negative	

Table B.2: 2D NARX model test results 16-40

Test no.	Mamm. no.	Real condition	Result	Pred. type.
16	91	Benign	True positive	Benign
17	92	Malign	True positive	Malign
18	93	Normal	True negative	
19	94	Normal	True negative	
20	95	Malign	True positive	Malign
21	96	Benign	True positive	Benign
22	97	Benign	True positive	Benign
23	98	Normal	True negative	
24	103	Normal	True negative	
25	111	Malign	True positive	Malign
26	117	Malign	True positive	Benign
27	119	Normal	True negative	
28	120	Malign	False negative	
29	131	Normal	False positive	
30	132	Benign	True positive	
31	133	Normal	True negative	
32	134	Malign	True positive	Malign
33	135	Normal	True negative	
34	136	Normal	True negative	
35	139	Normal	True negative	
36	140	Normal	True negative	
37	141	Malign	True positive	Malign
38	143	Normal	True negative	
39	144	Malign	True positive	Malign
40	150	Benign	True positive	Benign

Table B.3: 2D NARX model test results 41-64

Test no.	Mamm. no.	Real condition	Result	Pred. type.
41	151	Normal	True negative	
42	153	Normal	True negative	
43	154	Normal	True negative	
44	155	Malign	True positive	Malign
45	156	Normal	True negative	
46	158	Malign	True positive	Malign
47	160	Benign	True positive	Benign
48	166	Normal	True negative	
49	167	Benign	True positive	Benign
50	168	Normal	True negative	
51	169	Normal	True negative	
52	173	Normal	False positive	
53	174	Normal	True negative	
54	181-A	Malign	True positive	Malign
55	181-B	Malign	True positive	Malign
56	183	Normal	False positive	
57	184	Malign	True positive	Malign
58	186	Malign	True positive	Malign
59	189	Normal	True negative	
60	190	Benign	True positive	Benign
61	195	Benign	True positive	Benign
62	196	Normal	True negative	
63	202	Malign	True positive	Malign
64	203	Normal	True negative	

Table B.4: 2D NARX model test results 65-86

Test no.	Mamm. no.	Real condition	Result	Pred. type.
65	204	Benign	True positive	Benign
66	206-A	Malign	True positive	
67	206-B	Malign	True positive	Malign
68	207	Malign	True positive	Malign
69	209	Malign	True positive	Malign
70	211	Malign	True positive	Malign
71	212	Normal	True negative	
72	213	Malign	True positive	Malign
73	231	Malign	False negative	
74	232	Normal	True negative	
75	214	Normal	True negative	
76	218	Benign	True positive	Benign
77	234	Normal	True negative	
78	237	Normal	True negative	
79	238	Malign	True positive	Malign
80	246	Normal	True negative	
81	247	Normal	True negative	
82	248	Benign	False negative	
83	249	Malign	True positive	Malign
84	251	Normal	False positive	
85	255	Normal	True negative	
86	272	Normal	True negative	

Table B.5: 2D NARX model test results 87-100

Test no.	Mamm. no.	Real condition	Result	Pred. type.
87	273	Normal	True negative	
88	314	Benign	True positive	Benign
89	293	Normal	True negative	
90	294	Normal	True negative	
91	297	Normal	True negative	
92	299	Normal	True negative	
93	300	Normal	True negative	
94	301	Normal	True negative	
95	302	Normal	False positive	
96	303	Normal	True negative	
97	304	Normal	True negative	
98	305	Normal	True negative	
99	306	Normal	True negative	
100	314	Benign	True positive	Benign

Appendix C

MSRBF testing results: Tissue-type ratio 1/4

The appendix shows results of the first test (out of 4) made with different tissue compositions. Test 1 had 31.68% fatty, 29.20% dense and 38.9% glandular images.

Table C.1: MSRBF: Test tissue-type ratio 1/4, Results 1-9

test no.	Mamm.	Tissue	Class	Predict.	Result
1	**6**	FATTY	negative	negative	TN
2	**11**	FATTY	negative	negative	TN
3	**14**	GLAND	negative	negative	TN
4	**15**	GLAND	benign	negative	FN
5	**28**	GLAND	malign	malign, benign	TP
6	**29**	GLAND	negative	negative	TN
7	**31**	FATTY	negative	negative	TN
8	**32**	GLAND	benign	benign	TP
9	**41**	GLAND	negative	negative	TN

Table C.2: MSRBF: Test tissue-type ratio 1/4, Results 10-35

test no.	Mamm.	Tissue	Class	Predict.	Result
10	**49**	GLAND	negative	negative	TN
11	**60**	FATTY	negative	negative	TN
12	**61**	DENSE	malign	malign	TP
13	**63**	DENSE	benign	benign	TP
14	**64**	DENSE	negative	negative	TN
15	**67**	DENSE	negative	negative	TN
16	**68**	DENSE	negative	negative	TN
17	**69**	FATTY	3 benign	3 benign	TP
18	**71**	GLAND	negative	negative	TN
19	**74**	GLAND	negative	negative	TN
20	**75**	FATTY	malign	malign	TP
21	**77**	FATTY	negative	negative	TN
22	**78**	FATTY	negative	negative	TN
23	**84**	GLAND	negative	negative	TN
24	**86**	GLAND	negative	negative	TN
25	**88**	FATTY	negative	negative	TN
26	**89**	GLAND	negative	negative	TN
27	**93**	GLAND	negative	negative	TN
28	**98**	FATTY	negative	negative	TN
29	**100**	DENSE	negative	negative	TN
30	**104**	DENSE	benign	3 benign	TP
31	**107**	DENSE	benign	benign, malign	TP
32	**108**	DENSE	negative	negative	TN
33	**109**	DENSE	negative	negative	TN
34	**111**	DENSE	malign	benign, malign	TP
35	**112**	DENSE	negative	negative	TN

Table C.3: MSRBF: Test tissue-type ratio 1/4, Results 36-61

test no.	Mamm.	Tissue	Class	Predict.	Result
36	**118**	GLAND	negative	negative	TN
37	**119**	GLAND	benign	benign	TP
38	**120**	GLAND	malign	benign, malign	TP
39	**121**	GLAND	benign	benign	TP
40	**125**	DENSE	malign	malign	TP
41	**137**	DENSE	negative	negative	TN
42	**141**	FATTY	malign	negative	FN
43	**151**	FATTY	negative	negative	TN
44	**154**	FATTY	negative	negative	TN
45	**157**	FATTY	negative	negative	TN
46	**160**	FATTY	benign	benign	TP
47	**161**	GLAND	negative	negative	TN
48	**165**	GLAND	benign	malign	TP
49	**166**	GLAND	negative	negative	TN
50	**168**	FATTY	negative	negative	TN
51	**171**	GLAND	malign	malign	TP
52	**176**	GLAND	negative	negative	TN
53	**177**	GLAND	negative	negative	TN
54	**181**	GLAND	malign	benign	TP
55	**182**	GLAND	negative	negative	TN
56	**186**	GLAND	malign	negative	FN
57	**187**	GLAND	negative	negative	TN
58	**191**	GLAND	benign	negative	FN
59	**192**	GLAND	negative	negative	TN
60	**194**	DENSE	negative	negative	TN
61	**195**	FATTY	benign	negative	FN

Table C.4: MSRBF: Test tissue-type ratio 1/4, Results 62-87

test no.	Mamm.	Tissue	Class	Predict.	Result
62	**198**	DENSE	benign	benign, malign	TP
63	**200**	DENSE	negative	negative	TN
64	**202**	DENSE	malign	malign	TP
65	**205**	FATTY	negative	negative	TN
66	**208**	DENSE	negative	negative	TN
67	**210**	GLAND	negative	negative	TN
68	**211**	GLAND	malign	benign	TP
69	**213**	GLAND	malign	benign	TP
70	**218**	GLAND	benign	benign	TP
71	**219**	GLAND	benign	benign	TP
72	**220**	GLAND	negative	negative	TN
73	**224**	DENSE	negative	negative	TN
74	**228**	GLAND	negative	negative	TN
75	**230**	FATTY	negative	negative	TN
76	**231**	FATTY	malign	malign	TP
77	**233**	GLAND	malign	malign, benign	TP
78	**234**	GLAND	negative	negative	TN
79	**238**	FATTY	malign	malign	TP
80	**244**	DENSE	benign	benign, malign	TP
81	**248**	FATTY	benign	malign	TP
82	**249**	DENSE	malign	malign	TP
83	**250**	DENSE	negative	negative	TN
84	**251**	FATTY	negative	negative	TN
85	**252**	FATTY	benign	benign	TP
86	**254**	DENSE	negative	negative	TN
87	**255**	FATTY	negative	negative	TN

Table C.5: MSRBF: Test tissue-type ratio 1/4, Results 88-113

test no.	Mamm.	Tissue	Class	Predict.	Result
88	**256**	FATTY	malign	negative	FN
89	**258**	DENSE	negative	negative	TN
90	**261**	DENSE	negative	negative	TN
91	**265**	GLAND	malign	malign	TP
92	**267**	FATTY	malign	malign	TP
93	**269**	GLAND	negative	negative	TN
94	**270**	GLAND	malign	malign	TP
95	**271**	FATTY	malign	malign	TP
96	**272**	FATTY	negative	negative	TN
97	**275**	GLAND	negative	negative	TN
98	**278**	GLAND	negative	negative	TN
99	**282**	DENSE	negative	negative	TN
100	**287**	DENSE	negative	negative	TN
101	**289**	DENSE	negative	negative	TN
102	**292**	GLAND	negative	negative	TN
103	**293**	FATTY	negative	negative	TN
104	**295**	DENSE	negative	negative	TN
105	**300**	FATTY	negative	negative	TN
106	**302**	FATTY	negative	negative	TN
107	**305**	FATTY	negative	negative	TN
108	**307**	FATTY	negative	negative	TN
109	**309**	FATTY	negative	negative	TN
110	**314**	FATTY	2 benign	benign	TN
111	**317**	DENSE	negative	negative	TN
112	**319**	DENSE	negative	negative	TN
113	**322**	DENSE	negative	benign	FP