Computational intelligence & modeling of crop disease data in Africa

Owomugisha, Godliver

DOI:
10.33612/diss.130773079

Publication date:
2020

Computational intelligence & modeling of crop disease data in Africa

Godliver Owomugisha

This research was supported by College of Computing and Information Sciences, Makerere University with funding from the Bill and Melinda Gates Foundation (BMGF) project number OPP1112548

Computational intelligence & modeling of crop disease data in Africa
Godliver Owomugisha
ISBN: 978-94-034-2637-2 (printed version)
ISBN: 978-94-034-2636-5 (electronic version)

UNIVERSITY OF GRONINGEN

Computational intelligence & modeling of crop disease data in Africa

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus, Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Friday 28 August 2020 at 16.15 hours

by

Godliver Owomugisha

born on 2 August 1987
in Bushenyi, Uganda

Supervisors
Prof. M. Biehl
Prof. N. Petkov

Co-supervisors
Dr. E. Mwebaze
Dr. J.A. Quinn

Assessment committee
Prof. D. Karastoyanova
Prof. L.C. Verbrugge
Prof. B. Hammer

Contents

Acknowledgements **ix**

List of Figures **xi**

List of Tables **xiii**

1 Introduction **1**
1.1 Scope of this thesis 3
1.2 Outline 4

2 Learning Vector Quantization **7**
2.1 Introduction 8
2.2 Learning Vector Quantization and its variants 9
2.2.1 Classical LVQ 9
2.2.2 Generalized LVQ 10
2.2.3 Generalized Matrix LVQ 10

I Disease Diagnosis with Leaf Images **13**

3 Disease Incidence and Severity Measurements from Leaf Images **15**
3.1 Introduction 16
3.2 The Leaf Image Data 18
3.2.1 Disease leaf symptoms 19
3.3 Methods and experiments 20
3.3.1 Feature extraction 20
3.3.2 Classification of Disease Incidence 22

3.3.3 Classification of disease severity 24
3.4 System Deployment 25
3.5 Discussion 26

II Disease Diagnosis with Spectral Data **29**

4 Machine Learning for diagnosis of disease in plants using spectral data **31**
4.1 Introduction 32
4.2 Materials & Methods 34
4.2.1 Data collection 34
4.2.2 Image data processing 35
4.2.3 Spectral data pre-processing 36
4.2.4 Training a diagnosis classifier 39
4.3 Results 42
4.3.1 *Good* vs. *bad* part of leaves in spectral data 42
4.3.2 Image-based features vs spectral data 43
4.3.3 PCA spectral features 43
4.4 Discussion 44
4.5 Conclusion 45

5 Matrix relevance learning for multi-class classification with spectral data **47**
5.1 Introduction 48
5.2 The GMLVQ machine learning framework 51
5.2.1 Dimensionality reduction 52
5.3 Experiments 53
5.3.1 Experiment design and data collection 53
5.3.2 Data pre-processing 55
5.3.3 Training and validation 55
5.4 Results 56
5.4.1 Full spectral data 56
5.4.2 Reduced feature space 58
5.5 Discussion 61

6 Early detection of plant diseases using spectral data **63**
6.1 Introduction 64
6.2 Related work 65
6.3 Materials and methods 66
6.3.1 Experimental design and data collection 66
6.3.2 Confirmation of CBSD transmission 67

6.3.3 Data pre-processing and feature extraction 68
6.3.4 Dimensionality reduction 69
6.3.5 Prototype-based disease classification 70
6.3.6 Training and validation 71
6.4 Results 72
6.5 Discussion and Outlook 74

7 A low-cost 3-D printed smartphone add-on spectrometer for diagnosis of crop diseases in the field 77
7.1 Introduction 78
7.2 Materials and methods 79
7.2.1 Commercial spectrometer 79
7.2.2 Customised spectrometer design 80
7.2.3 Methods 81
7.2.4 Data pre-processing 83
7.3 Results 85
7.4 Discussion 86

8 Summary and Outlook 87
8.1 Future work 88

Bibliography 90
8.2 Toekomstwerk 103

Acknowledgments

This PhD journey has been a learning, challenging and interesting part of my life and I'm so grateful to everybody that encouraged, inspired and supported me to follow this dream.

First, I express my sincere gratitude to my main supervisors: (i). Prof. Michael Biehl for his tireless effort, patience, motivation and continuous support during my study. Besides the academic work, I will not forget to thank him for the delicious dinners he prepared each year for his Ph.D students. It was always wonderful and stress relieving I must say. (ii). Dr. Ernest Mwebaze, I'm so thankful for the scholarship opportunity I got through your grant from Bill and Melinda Gates Foundation. Even when I was given a Ph.D offer on your project, I kept asking myself questions if I would work according to your expectations but your patience, guidance and continuous support kept me going and I'm glad for this achievement. (iii). Dr. John A. Quinn, to this level, I'm so thankful for the mentorship you gave me. Many times I think, I would be lost in another research discipline. Your brilliant ideas for research in developing countries and motivation and career guidance right from the time I was your masters student have brought up this achievement. I will not forget to thank you and your lovely wife Sofie for the lively BBQ's the AI-research team has had at your lovely home in Namulonge.

Special thanks go to Prof. Udo Seiffert and Friedrich Melchert from Fraunhofer Institute for Factory Operation and Automation IF, Magdeburg/Germany, for the collaboration we have had from you on this research project. In the same spirit, I thank our collaborators at NACRII especially, Dr. Christopher Omongo, Dr. Ephraim Nuwanamya and Dalton Kanyesigye.

My heartfelt thanks also goes to the Intelligent Systems research group at the University of Groningen. I thank Prof. Nicolai Petkov, the head of the group who accepted me to join and do research under this group. The group dinners you organized for us will always be memorable. I would also like to thank Prof. Michael Wilkinson for warm discussions during our lunch breaks. Great thanks also goes to: Kerstin, George, Nichola, Estefania, Ahmed, Jiapan, Laura Fernandez, Laura Fiorini, Maria, Rick, Aleke, Xiaoxuan, Wang, Caroline, Sreejita, M. Muhammedi, M. Babai, Astone, Simon, Hyoyin, Swarloop and Abol. I have very fond memories of my stay at the department.

Similar profound gratitude goes to the Artificial Intelligence & Data Science Research Group at Makerere University. The team I worked with: Dr. Joyce Nakatumba-Nabende, Pius, Daniel Mutembesa, Barbara, Solomon, Flavia, Jeremy, Benjamin, Lilian, Eugien, Rose Nakasi, Rose Nakibule, Martine Mubangizi, Samiha, Hewitt, Pamela, Daniel Ssendiwala, Claire, Gloria, Ali and all our interns.

I am also appreciative to my employer Busitema University for the support you have rendered to me to this level. Joining the University as a Bachelors holder, I thank you for the many recommendations and wonderful opportunities that have come along my way as your employee.

To my friends in Groningen, Elfie, Shrin, Nadia, Ertha, your lovely Mom Mrs. Liz and Carien. You have made my life outside the school campus so lively. I thank God meeting kind people like you.

To my lovely family: My lovely husband Bajurizi Tomson, you are such a gift sent by God. Thank you so much for standing by me and supporting me to follow my dream. Above all, I thank you for being there for our children Jordan Woods Biganja and Ann Kristal Woods especially in times of my study trips. To my sister Doreen and your husband, thank you for giving us a supporting hand to take care of our children in times work schedules got so difficult on us. And to your lovely children Mariah and Joseph. I also thank my big cousin Kiconco Sylivia, a sister, a friend and a counsellor. For times when life went astray, you gave a listening ear and made sure life gets back to the right place. Good luck in your Ph.D journey as well. Special thanks goes to my Dad Mugisha Grandford. You are my number one hero in this world! Raising the four of us for twenty (20) years as a single dad after the loss of our beloved mother was a very big sacrifice. May the dear good Lord reward you abundantly. To our lovely new mom, little brothers John and Simon, Christine, may God bless you to see the great heights. Lastly, I thank my big brothers Tumuhaise Grandford and Tugume Godfrey for keeping the family one unit.

Godliver Owomugisha
Groningen
June 11, 2020

List of Figures

3.1 Experts assessing plants & scoring diseases in the field 17
3.2 Sample images associated with the five disease classes of the classification problem. 18
3.3 Examples of histograms (bottom) extracted from the corresponding healthy and diseased images (top). 21
3.4 Image with ORB interest keypoints identified 22
3.5 Sample images associated with the five severity levels for CMD (top) and CBSD (bottom). 24
3.6 Screenshots of the smartphone application for remote diagnosis of crop health. 26

4.1 Crop effect as a result of late diagnosis (a - clean cassava tuber; b,c,d - severe effects caused by CBSD disease) 32
4.2 Cassava disease automated diagnostic pipeline as described in 4.2 35
4.3 Data collection in the field & depiction of good and bad part of leaf . 36
4.4 Spectral data in raw form, illustrating mean spectra (over classes). We consider the region between 400 nm $-$ 900nm after truncating the smallest and largest wavelengths marked by the vertical lines. 40
4.5 Overall accuracy (%) with increasing number of principal components with GMLVQ algorithm. 44

5.1 Depiction of asymptomatic(good) and symptomatic(bad) part of a leaf . . . 54
5.2 Example images of leaves of cassava manifesting the different diseases. 55
5.3 Illustration for class-conditional means of Cassava spectral data not individual spectra. The left panel displays raw, full signal, the right panel shows the corresponding pre-processed spectra. 55

5.4 Feature relevance as quantified by diagonal elements of Λ, cf. Eq. (5.3), for original spectra as feature vectors. 57
5.5 Visualization of the dataset depicting the three major classes in the dataset plotted as projections of feature vectors (original spectra) on the two leading eigenvectors of GMLVQ relevance matrix. 58
5.6 Visualization of GMLVQ prototypes of the original spectra 59
5.7 Performance of classifiers based on N Principal Component (left) and n coefficients in the polynomial representation (right panel). 59
5.8 Selection of features with diagonal relevances (GMLVQ) above a threshold. . 60
5.9 Diagonal relevances of GMLVQ in original feature space as reconstructed after performing the training in terms of 30 (left panel) and 5 (right panel) principal components. 61
5.10 Receiver operating characteristic curves for one class vs All (multi-class problem). Top-left panel shows CBSD vs All and CMD vs All in the original feature space (400 - 900nm. Top-right panel shows CBSD vs All and CMD vs All with reduced features (peak selection). The bottom panel shows Healthy vs All both in the original space and reduced features (peak selection). The solid lines refer to AUC in original feature space (400 - 900nm) while the dashed lines refer to AUC with peak selection between 500 - 600nm. 62

6.1 Sample cassava crops grown in the screen house setting. 67
6.2 Spectral data in original form. Mean spectra of healthy samples and diseased samples are shown, respectively. 69
6.3 Feature relevance as quantified by diagonal elements of Λ, cf. Eq. (2) (left), feature representation in the coefficient space with PCA (right). In chapter 5, we explain the feature selection process where spectral bands 500 - 600 nm were found to be more relevant 72
6.4 The top-left graph illustrates the ground truth in terms of virus load based on RT-PCR analysis. The top-right and the bottom panels display GMLVQ scores S, Eq. (6.5), for individual plants (top-right) and on average over classes (bottom). The top-right panel corresponds to the original space with wavelengths 500-600 nm. The bottom graph shows results of combining GMLVQ with PCA with 30 coefficients. . 74
6.5 Class-wise training error in original space (left) and PCA (right) . . . 75
6.6 Receiver operating characteristic curves for Healthy vs CBSD with GMLVQ algorithm in the original space of the spectrum and in the coefficients with PCA 75

7.1 Diffraction grating. The numbers near the screen are n values, the order of the image. Taken from (Burchill 2019). 80
7.2 Architectural design 81
7.3 First prototype 82
7.4 Adapter design for the 3D-printed smartphone case. Actual designs are available at https://github.com/godliver/3-D-Printouts.git. . . . 82
7.5 Spectral data in an image array form acquired with the setup in Figure 7.3 83
7.6 Color histograms, a transformation from color RGB spectra in Figure 7.5 83
7.7 Corresponding spectral data acquired with Aspectra mini application. 84
7.8 Projection on eigenvectors of Principal components. On left panel are data points of color histograms. The right panel are data points acquired by the Aspectra mini application. 84

List of Tables

3.1 Overall 10-fold cross-validated accuracy scores (%) for different algorithms applied to the different leaf image representations. 23
3.2 Confusion matrix for healthy class vs. four major diseases 25
3.3 Confusion matrix for combined severity 25

4.1 Spectral data dependence on leaf quality (Healthy vs. CBSD, CMD). Overall cross-validation accuracy scores (%) for the different algorithms. 42
4.2 Confusion matrix (%) for *Bad* part of leaf with GMLVQ 43
4.3 Confusion matrix (%) for *Good* part of leaf with GMLVQ 43
4.4 Overall cross-validation accuracy scores (%) with different data features (Healthy vs. CBSD, CMD) 44

5.1 Overall accuracy in original feature space (400 - 900 nm) and when applying dimensional reduction techniques. 60

6.1 Accuracy in original feature space vs. dimensional reduction obtained on average over validation. (a) using original data full spectra between 400 - 900 nm and (b) using original data between 500 - 600 nm. In the coefficient space we use 30 dimensions for all algorithms. 73

7.1 Overall accuracy score Aspectra Mini vs. Color Histograms 85
7.2 Confusion matrix for Aspectra Mini with Extra trees 86
7.3 Confusion matrix for Color Histograms with Extra trees 86