The Second Australian Workshop on Artificial Intelligence in Health AIH 2012

held in conjunction with the

25th Australasian Joint Conference on Artificial Intelligence (AI 2012)

Tuesday, 4th December 2012
Sydney Harbour Marriott Hotel, Sydney, Australia

WORKSHOP PROCEEDINGS

Editors : Sankalp Khanna, Abdul Sattar, David Hansen

ACKNOWLEDGEMENTS

Program Chairs

- Abdul Sattar (Griffith University, Australia)
- David Hansen (CSIRO Australian e-Health Research Centre, Australia)

Workshop Chair

- Sankalp Khanna (CSIRO Australian e-Health Research Centre, Australia)

Senior Program Committee

- Aditya Ghose (University of Newcastle, Australia)
- Anthony Maeder (University of Western Sydney, Australia)
- Wayne Wobcke (University of New South Wales, Australia)
- Mehmet Orgun (Macquarie University, Australia)
- Yogesan (Yogi) Kanagasingam (CSIRO Australian e-Health Research Centre, Australia)

Program Committee

- Simon McBride (CSIRO Australian e-Health Research Centre)
- Adam Dunn (University of New South Wales)
- Stephen Anthony (University of New South Wales)
- Lawrence Cavedon (Royal Melbourne Institute of Technology / NICTA)
- Diego Mollá Aliod (Macquarie University)
- Michael Lawley (CSIRO Australian e-Health Research Centre)
- Anthony Nguyen (CSIRO Australian e-Health Research Centre)
- Amol Wagholikar (CSIRO Australian e-Health Research Centre)
- Bevan Koopman (CSIRO Australian e-Health Research Centre)
- Kewen Wang (Griffith University)
- Vladimir Estivill-Castro (Griffith University)
- John Thornton (Griffith University)
- Bela Stantic (Griffith University)
- Byeong-Ho Kang (University of Tasmania)
- Justin Boyle (CSIRO Australian e-Health Research Centre)
- Guido Zuccon (CSIRO Australian e-Health Research Centre)
- Hugo Leroux(CSIRO Australian e-Health Research Centre)
- Alejandro Metke (CSIRO Australian e-Health Research Centre)

Key Sponsors

- CSIRO Australian e-Health Research Centre
- Institute for Integrated and Intelligent Systems, Griffith University

Supporting Organisations

- The Australasian College of Health Informatics
- The Australasian Medical Journal
- The Australasian Telehealth Society

PROGRAM

8:30 am – 9:00 am	Registration and Welcome
9:00 am – 10:30 am	Session 1 — Chair : Abdul Sattar Keynote Address Technology in Healthcare: Myths and Realities *Dr. Jia-Yee Lee* *National ICT Australia (NICTA)* Keynote Address Driving Digital Productivity in Australian Health Services *Dr. Sankalp Khanna* *CSIRO Australian e-Health Research Centre*
10:30 am – 11:00 am	Morning Tea
11:00 am – 12:30 pm	Session 2 — Chair : Sadananda Ramakoti An investigation into the types of drug related problems that can and cannot be identified by commercial medication review software *Colin Curtain, Ivan Bindoff, Juanita Westbury and Gregory Peterson* FS-XCS vs. GRD-XCS: An analysis using high-dimensional DNA microarray gene expression data sets *Mani Abedini, Michael Kirley and Raymond Chiong* Reliable Epileptic Seizure Detection Using an Improved Wavelet Neural Network *Zarita Zainuddin, Pauline Ong and Kee Huong Lai* Clinician-Driven Automated Classification of Limb Fractures from Free-Text Radiology Reports *Amol Wagholikar, Guido Zuccon, Anthony Nguyen, Kevin Chu, Shane Martin, Kim Lai and Jaimi Greenslade* Using Prediction to Improve Elective Surgery Scheduling *Zahra Shahabi Kargar, Sankalp Khanna and Abdul Sattar*
12:30 pm – 2:00 pm	LUNCH (and Poster Session)
2:00 pm – 3:30 pm	Session 3 — Chair : Wayne Wobcke Acute Ischemic Stroke Prediction from Physiological Time Series Patterns *Qing Zhang, Yang Xie, Pengjie Ye and Chaoyi Pang* Comparing Data Mining with Ensemble Classification of Breast Cancer Masses in Digital Mammograms *Shima Ghassem Pour, Peter Mc Leod, Brijesh Verma and Anthony Maeder* Automatic Classification of Cancer Notifiable Death Certificates *Luke Butt, Guido Zuccon, Anthony Nguyen, Anton Bergheim and Narelle Grayson* If you fire together, you wire together; Hebb's Law revisited *Prajni Sadananda and Sadananda Ramakoti*
3:30 pm – 4:00 pm	Afternoon Tea
4:00 pm – 5:30 pm	Session 4 — Chair : Sankalp Khanna Keynote Address Smart Analytics in Health *Dr. Christian Guttman* *IBM Research Australia* Panel Discussion AI in Health : the 3 Big Challenges *Panel Chair : Professor Abdul Sattar . Panelists : Dr. Jia-Yee Lee, Dr. Christian Guttman, Prof. Wayne Wobcke, Prof. Sadananda Ramakoti*
5:30 pm	Announcement of Best Paper Award Workshop Close

TABLE OF CONTENTS

Second Australian Workshop on Artificial Intelligence in Health (AIH 2012)

PREFACE

Sankalp Khanna[1, 2], Abdul Sattar[2], David Hansen[1]

[1] The Australian e-Health Research Centre, RBWH, Herston, Australia
{Sankalp.Khanna, David.Hansen}@csiro.au
[2] Institute for Integrated and Intelligent Systems, Griffith University, Australia
A.Sattar@griffith.edu.au

1 Motivation behind the workshop series

The business of health service delivery is a complex one. Employing over 850,000 people, and delivering services to 21.3 million residents, the Australian healthcare system is currently struggling to deal with increasing demand for services, and an acute shortage of skilled professionals. The National e-Health Strategy drives a nationwide agenda to provide the infrastructure and tools required to support the planning, management and delivery of health care services. National initiatives such as the National Health Reform Program, the National Broadband Network, and the Personally Controlled Electronic Health Record are accelerating the use of information and communication technologies in delivering healthcare services. The Australasian Joint Conferences on Artificial Intelligence (AI) provide an excellent opportunity to bring together artificial intelligence researchers who are working in health research.

Driven by a senior program committee comprising distinguished faculty from several Australian universities including Griffith University, the University of New South Wales, University of Newcastle, University of Western Sydney, and Macquarie University, and specialist health research organisations including the CSIRO Australian e-Health Research Centre, the Artificial Intelligence in Health workshop series was created in 2011 to bring these researchers together as part of Australia's premier Artificial Intelligence conference.

2 AIH 2011 – the First Australian Workshop on Artificial Intelligence

Held for the first time in December 2011, the workshop was the first of its kind to bring together scholars and practitioners nationally in the field of Artificial Intelligence driven Health Informatics to present and discuss their research, share their knowledge and experiences, define key research challenges and explore possible collaborations to advance e-Health development nationally and internationally. The workshop was co-located with the 24th Australasian Joint Conference on Artificial Intelligence and was attended by 25 delegates.

Of the 16 submissions received, 6 were accepted as Full Papers and 5 as Short Papers accompanied with posters. All papers presented at the AIH 2011 workshop were also invited to revise and submit their manuscripts for inclusion in a special issue of the Australasian Medical Journal. Of these, seven papers and a letter to the editor were published in the special issue in September 2012.

3 AIH2012 - The Second Australian Workshop on Artificial Intelligence

The Second Australian Workshop on Artificial Intelligence (AIH 2012) is being held in conjunction with the 25th Australasian Joint Conference on Artificial Intelligence (AI 2012) in Sydney, Australia, on the 4th of December, 2012.
The Call for Papers received an excellent response this year. All submitted papers went through a rigorous review process. Of these, 6 full papers and 3 short papers have been accepted for presentation in the workshop and for publication in these CEUR proceedings. The workshop will also feature three keynote addresses and a panel discussion on the topic “AI in Health: the 3 Big Challenges”.

This year again, the workshop is offering 4 travel scholarships of $250 each to students who were first authors of accepted papers. A best paper prize of $250 will also be awarded on the workshop day. Both prizes have been sponsored by the CSIRO Australian e-Health Research Centre.
All accepted full papers and short papers will be also invited to extend and reformat their papers for publication in a special issue of the Australasian Medical Journal (www.amj.net.au). The journal is indexed on the following databases: DOAJ, EBSCO, Genamics journalseek, ProQuest, Index Copernicus,

Open J-Gate, Intute, Global health and CAB Abstracts databases, MedWorm, Scopus, Socolar, PMC, PubMed.

4 Workshop Organisation

4.1 Program Chairs

Abdul Sattar (Griffith University, Australia)
David Hansen (CSIRO Australian e-Health Research Centre, Australia)

4.2 Workshop Chair

Sankalp Khanna (CSIRO Australian e-Health Research Centre, Australia)

4.3 Senior Program Committee

Aditya Ghose (University of Newcastle, Australia)
Anthony Maeder (University of Western Sydney, Australia)
Wayne Wobcke (University of New South Wales, Australia)
Mehmet Orgun (Macquarie University, Australia)
Yogesan (Yogi) Kanagasingam (CSIRO Australian e-Health Research Centre, Australia)

4.4 Program Committee

Simon McBride (CSIRO Australian e-Health Research Centre)
Adam Dunn (University of New South Wales)
Stephen Anthony (University of New South Wales)
Lawrence Cavedon (Royal Melbourne Institute of Technology / NICTA)
Diego Mollá Aliod (Macquarie University)
Michael Lawley (CSIRO Australian e-Health Research Centre)
Anthony Nguyen (CSIRO Australian e-Health Research Centre)
Amol Wagholikar (CSIRO Australian e-Health Research Centre)
Bevan Koopman (CSIRO Australian e-Health Research Centre)
Kewen Wang (Griffith University)
Vladimir Estivill-Castro (Griffith University)
John Thornton (Griffith University)
Bela Stantic (Griffith University)

Byeong-Ho Kang (University of Tasmania)
Justin Boyle (CSIRO Australian e-Health Research Centre)
Guido Zuccon (CSIRO Australian e-Health Research Centre)
Hugo Leroux(CSIRO Australian e-Health Research Centre)
Alejandro Metke (CSIRO Australian e-Health Research Centre)

4.5 Key Sponsors

CSIRO Australian e-Health Research Centre
Institute for Integrated and Intelligent Systems, Griffith University

4.6 Supporting Organisations

The Australasian College of Health Informatics
The Australasian Medical Journal
The Australasian Telehealth Society

5 Acknowledgements

We are especially thankful to the organising committee of the 25th Australasian Joint Conference on Artificial Intelligence (AI 2012). This workshop series would not have possible without their support. We would also like to thank the Workshop Chair of AI 2012, Hans Guesgen, for organising the workshops and championing these CEUR workshop proceedings.

Technology in Healthcare : Myths and Realities

Keynote Address

Dr. Jia-Yee Lee

National Information and Communications Technology Australia Ltd (NICTA), Australia
jia-yee.lee@nicta.com.au

Speaker Profile

Dr Jia-Yee Lee is the Director of the Health and Life Science Business Team at National ICT Australia Ltd. She manages the business, commercial and research activities of the NICTA groups in Diagnostic and Computational Genomics, Biomedical Informatics, Portable Motion Analytics and Bio-Imaging Analytics. Prior to joining NICTA, Jia-Yee spent 10 years in the management consulting sector providing leadership in developing and implementing strategies and operational plans that improved business outcomes for clients in government, ICT, and healthcare sectors. Her business plans have led to international and national investments into Australian-based start-ups. Jia-Yee has extensive experience as a project manager working on complex multi-disciplinary and multi-million dollar programs funded by State and Commonwealth governments. Her e-health experience includes stakeholder engagement with clinicians and leading technical teams to implement a range of commercial web-based systems for the healthcare and medical research sectors. With more than 20 years in medical research, Jia-Yee has led programs at MacFarlane Burnet Centre (now "Burnet Institute") and the Victorian Infectious Diseases Reference Laboratory, Melbourne Health. Her research into hepatitis B virus and rubella virus was funded by the National Health and Medical Research Council of Australia. Jia-Yee's research skills include molecular and diagnostic virology, and electron and confocal microscopy. Jia-Yee has a PhD from the University of Melbourne and a MBA from Melbourne Business School.

Driving Digital Productivity in Australian Health Services

Keynote Address

Sankalp Khanna

The Australian e-Health Research Centre, RBWH, Herston, Australia
Sankalp.Khanna@csiro.au

Speaker Profile

Sankalp is a Postdoctoral Fellow at the Australian e-Health Research Centre, the leading national research facility applying information and communication technology to improve health services and clinical treatment for Australians. As a member of the Forecasting and Scheduling team, he is actively engaged in projects in the areas of planning and optimization, patient flow analytics, prediction and forecasting, and predictive scheduling, all aimed at employing artificial intelligence to improve the efficiency of the health system.

His research interests include Applied Artificial Intelligence, Prediction and Forecasting, Planning and Scheduling, Multi Agent Systems, Distributed Constraint Reasoning, and Decision Making and Learning under Uncertainty.

Sankalp completed a PhD in 2010 looking at intelligent techniques to model and optimise the complex, dynamic and distributed processes of Elective Surgery Scheduling. He was the recipient of a state award for outstanding student achievement in 2006. He has co-authored several journal and conference papers and editorials, and served on the program and organising committees of numerous national and international conferences and workshops. He is a member of the ACS, HISA, IEEE and AAAI societies.

Sankalp was also the founding workshop chair of this AI in Health workshop series.

Smart Analytics in Health

Keynote Address

Christian Guttman

IBM Research, Australia
Christian.guttmann@au1.ibm.com

Speaker Profile

Dr. Guttmann leads and defines projects around health care at the newly established IBM Research labs in Melbourne – the 11th lab of IBM Research worldwide. One focus of Guttmann's work is to build smarter analytics that enables health care entities (doctors, nurses, hospitals, pharmacies, etc) to collaborate more efficiently in complex environments. His work addresses the information and communication challenges faced by tomorrow's world of health care: How can we create and apply smarter collaborative health care technologies that cope with the tsunami of chronic diseases.

Prior to IBM, Dr. Guttmann led the research theme on health care and disaster at the Etisalat British Telecom Innovation Centre (EBTIC). The theme partnered with major stakeholders, including governmental health authorities and ministries. He has been a research fellow at the Faculty of Medicine, Nursing and Health Sciences at Monash University, where he researched how intelligent systems can improve collaborative care (done in together with primary health care providers). He worked also in industrial projects with HP and Ericsson.

Dr. Guttmann holds a PhD degree from Monash University, two Master degrees from Paderborn University (Germany) and the Royal Institute of Technology (Sweden), and a psychology degree from Stockholm University (Sweden). He organised major conferences and workshops, edited two books on intelligent agent technologies, and co-authored over 30 articles in leading conferences and journals.

An investigation into the types of drug related problems that can and cannot be identified by commercial medication review software

Colin Curtain, Ivan Bindoff, Juanita Westbury and Gregory Peterson

Unit for Medication Outcomes Research and Education
School of Pharmacy
University of Tasmania
{Colin.Curtain, Ivan.Bindoff, Juanita.Westbury, G.Peterson}@utas.edu.au

Abstract.
A commercially used expert system using multiple-classification ripple-down rules applied to the domain of pharmacist-conducted home medicines review was examined. The system was capable of detecting a wide range of potential drug-related problems. The system identified the same problems as pharmacists in many of the cases. Problems identified by pharmacists but not by the system may be related to missing information or information outside the domain model. Problems identified by the system but not by pharmacists may be associated with system consistency and perhaps human oversight or human selective prioritization. Problems identified by the system were considered relevant even though the system identified a larger number of problems than human counterparts.

Keywords: Clinical decision support system, multiple-classification ripple-down rules, expert system, pharmacy practice

1 Introduction

A drug-related problem (DRP) can be broadly defined as "…an event or circumstance involving drug therapy that actually or potentially interferes with desired health outcomes"[1] DRPs comprise a spectrum of problems including over- or under-dosage, drug-drug or drug-disease interactions, untreated disease and drug toxicity. Patient health education and compliance with therapy may be sub-standard and subsequently also be considered as drug-related problems. DRPs can be dangerous; For instance, a marginally high daily dose of warfarin has the potential to cause fatal bleeding.

Home medicines review (HMR) is a Commonwealth Government funded service conducted by accredited pharmacists to identify and address DRPs among eligible patients [2]. The main aims of the service are to enhance patient knowledge, quality use of medicines, reconcile health professional awareness of actual medication use and, ultimately, improve patient quality of life. The HMR service is a collaborative

activity between health professionals, typically accredited pharmacists, general practitioners (GPs), and patients. Since its inception in 2001 the service has steadily grown with nearly 80,000 HMRs funded in the 2011/2012 period [3].

An HMR is initiated for eligible consenting patients by a GP. Eligible patients are identified if they regularly take 5 or more medications among other criteria [2]. An HMR accredited pharmacist then obtains medical information from the GP, covering medical history, current medications and pathology.

A core component of an HMR is an interview between the pharmacist and the patient, with interview typically conducted in the patient's home. The interview, elicits additional information such as: actual medication use, additional non-prescribed medications, an understanding of the patient's motivation behind actual rather than directed medication use, and the patient's health and medication knowledge [4]. This process allows for a deeper understanding of the patient's situation and gives the pharmacist insight into cultural or language barriers, physical and economic limitations and family support.

The amassed information is reviewed by the pharmacist to identify actual and potential DRPs. The pharmacist writes a report of findings for the patient's GP, which includes recommendations to resolve any actual or potential problems. Consultation between the GP and the patient culminates in an actionable medication management plan designed to trial changes to existing therapy, and ideally, lead to improved medication use and improved patient health outcomes [4].

An important component is the professional skill of the pharmacist to be able to identify clinically relevant DRPs from the available information. This requires a wide scope of knowledge, not only of medications, but of evidence-based guidelines and contemporary management of a variety of medical conditions.

Evidence-based guidelines can be difficult to implement due to their apparent complexity. An example is provided from Basger *et al.'s Prescribing Indicators in Elderly Australians*: "Patient at high risk of a cardiovascular event (b) is taking an HMG-CoA reductase inhibitor (statin)"[5] If a patient did not meet this criterion this would be considered a DRP. It can be reasonably expected that pharmacists would be aware of statin medications currently available in Australia, in October 2012 these were: atorvastatin, fluvastatin, pravastatin, rosuvastatin, and simvastatin. Note (b) specifies those patients at high risk of cardiovascular event: "age>75 years, symptomatic cardiovascular disease (angina, MI[myocardial infarction], previous coronary revascularization procedure, heart failure, stroke, TIA[transient ischemic attack], PVD[peripheral vascular disease], genetic lipid disorder, diabetes and evidence of renal disease (microalbuminuria and/or proteinuria and/or GFR[glomerular filtration rate]<60ml/min". Determining patients at high risk of cardiovascular events is more problematic and requires sufficient additional information to make such a determination. One obvious problem is the amount of information that needs to be screened, both within the guideline text and the patient data, to identify appropriate patients.

A commercial product developed by Medscope, Medication Review Mentor (MRM)[6], incorporates a clinical decision support (CDSS) tool to assist with the detection of DRPs. MRM utilizes a knowledge-based system to detect DRPs and provide recommendations for their resolution. This knowledge-based system uses the

multiple classification ripple-down rules (MCRDR) method and was based on the work of Bindoff *et al.* who applied this approach to the knowledge domain of medication reviews [7, 8]. The ripple-down rules method was considered appropriate as knowledge could be gradually added to the knowledge base, broadening the scope and refining existing knowledge as the system was being used [7, 9]. Bindoff *et al.* suggested intelligent decision support software developed for this knowledge domain may improve the quality and consistency of medication reviews.

No prior research had been undertaken to determine the clinical decision support capacity of this commercial software, apart from contemporary research by the authors. This contemporary research by the authors assessed opinions from pharmacology experts and had determined that MRM is capable of identifying clinically relevant DRPs [10-12].

This evaluation attempts to provide light on the scope of DRPs that can be identified by this software by presenting summary counts and examples of the types of problems that were identified by MRM and by pharmacists. This paper evaluates the similarities and differences between pharmacist findings and MRM findings more in terms of a qualitative comparison by highlighting common findings, extremes of difference and discussing the possible advantages and limitations of the software, as well as discussing areas for potential improvements.

2 How MRM works

The decision support component of MRM is a knowledge-based system which uses MCRDR as its inference engine. MCRDR provides the knowledge engineer a way to incrementally improve the quality of the knowledge base through the addition of either new rules – which are added when the system fails to identify a DRP, or refinements to existing rules – which are added when the system incorrectly identifies an inappropriate DRP. The system's knowledge base is managed by medication review experts, who regularly review cases, examining the findings of the system for that case, and then adding/refining rules until the system produces a wholly correct set of findings for that case [8]. The validity of new rules is always being ensured, as the system identifies any conflicts which may arise from the addition of the new rule, and prompts the pharmacist to refine their rule until no further conflicts arise.

3 Methods

Australia-wide data collected during 2008 for a previous project, examining the economic value of HMRs, was used for this study [13]. The data contained patient demographics, medications, diagnoses and pathology results for 570 community-dwelling patients aged 65 years old and older. The 570 HMRs were obtained from 148 different pharmacists. Supplementing this data were the original reviewing pharmacists' findings, detailing pharmacist-identified DRPs and recommendations.

The HMR data were entered into MRM and DRPs identified by MRM were recorded. MRM utilized a wide range of information including basic patient de-

mographics such as age and gender, medication type including strength, directions and daily dose. MRM could calculate daily dose from strength and directions in many cases. Duration of use of medication could be entered, which included options of less than 3 months and more than 12 months. Medications were assigned Anatomic Therapeutic Chemical classifications (ATC) [14]. ATC is a five-tier hierarchical classification system allowing medications with similar properties to be grouped together in chemical classes which are then grouped into therapeutic categories.

Diagnoses could be entered and were based on the ICPC2 classifications [15]. The ICPC2 classification system was also hierarchical, grouping diagnoses under similar categories. Diagnoses could be assigned temporal context as recent, ongoing or past history. Medication allergies and general observations including height, weight and blood pressure could be entered. A wide range of pathology readings could be entered, including biochemical and hematological data.

At the time of the data entry and collections of results, August 2011, MRM contained approximately 1800 rules [16]. Rule development was undertaken by a pharmacist with expertise in both clinical pharmacology and HMRs [6].

Direct comparison of the DRPs identified by MRM and those identified by the original pharmacists was not possible due to the individual textual nature of each DRP. Each DRP identified by either the pharmacist or MRM was mapped to a concept (defined here as a theme) that described the DRP in sufficient detail to allow comparisons of similarity and difference between pharmacists and MRM. The themes often described the type of drug or disease and other relevant factors involved. The development of a list of themes and the mapping of DRPs to themes was performed manually by the author, a qualified pharmacist.

Examples of the text of two DRPs identified by a pharmacist and by MRM in the same patient are shown in Table 1. These DRPs were assigned the theme *Hyperlipidemia under/untreated*, which captured the basic problem identified within the text of each DRP.

Table 1. Example DRP text

MRM	**Pharmacist**
Patient has elevated triglycerides and is only taking a statin. Additional treatment, such as a fibrate, may be worth considering	Patient's cholesterol and triglycerides remain elevated despite Lipitor [statin]. This may be due to poor compliance or an inadequate dose

These themes provided a common language for comparison of the DRPs found by the original pharmacist reviewer and MRM. The initial themes were created where at least two of three published prescribing guidelines for the elderly [5, 17, 18] were in agreement concerning the same types of DRPs. DRPs from MRM and pharmacists were mapped to this table of themes. Further themes were added if both pharmacist and MRM DRPs could be mapped to any remaining 'non-agreement' prescribing guideline DRPs. New themes were developed for remaining pharmacist and MRM DRPs where concepts were clearly similar but were not contained within prescribing guidelines. These new themes were very broad such as *Vitamin, no indication*, and

may have included the DOCUMENT DRP classification text such as, *Therapeutic dose too high* [19]. The remaining DRPs were unique to either pharmacists or MRM and themes were provided where possible, such as, Skin disease *(un)dertreated* – pharmacist only DRP. Lastly miscellaneous otherwise unclassifiable DRPs were assigned *Other DRP pharmacist* and *Other DRP MRM*.

A list of 129 themes was developed. Many themes described disease states and/or drug classes describing identified DRPs in general terms. A descriptive analysis of the themes was performed.

The number of unique themes found in each patient was considered more important than the raw number of themes found in each patient. That is where two DRPs matched the same theme in the same patient, that theme was counted once. The reason behind this decision was to compare the number of different types of conceptual problems that could be identified across patients rather than raw numbers across patients.

Each theme identified in each patient was allocated into one of three categories: 1. Identified by pharmacists only, 2. Identified by MRM only or 3. Identified by both.

4 Results

The patient cohort was predominantly female, with an average age of 80 and an average of 12 medications and 9 diagnoses, as described in Table 2.

Table 2. Patient Demographics

Patient (N = 570)	**Demographics**
Age (years)	79.6 ± 6.7
Gender	Male 234 : Female 336
Number of medications	12.0 ± 4.4
Number of diagnoses	9.1 ± 5.2

Pharmacists identified a total of 2020 DRPs, an average of 3.5±1.8 per patient, with a range of 0 to 13 DRPs. MRM identified 3209 DRPs, of which 256 were excluded due to duplicated findings, leaving 2953 MRM DRPs, and an average of 5.2±2.8 per patient, ranging from 0 to 16 DRPs.

The 2953 MRM DRPs were able to be assigned to 100 different themes that described in general terms the central issue of each of the DRPs. Similarly, the 2020 pharmacist DRPs were able to be assigned to 119 different themes. Ninety of these themes which were identified by pharmacists were also able to be identified by MRM. Within these 90 themes, the software was able to identify the same issues as the pharmacists in one or more of the same patients for 68 particular themes.

The number of different themes identified by MRM or by pharmacists per patient was considered more important than the raw totals. The 2953 MRM DRPs were aggregated into 2854 themes. Pharmacist DRPs which were clearly identifiable as compliance or non-classifiable cost-related problems and outside the scope of MRM's

ability to identify were excluded, leaving 1726 pharmacist DRPs which were aggregated into 1680 themes.

MRM was able to identify the same themes as identified by pharmacists in the same patients 389 times, a 23% (389/1680) overlap of pharmacist findings by theme and patient. This then left 1291 themes identified by pharmacists only and 2465 themes identified by MRM only. For each patient a Jaccard coefficient was calculated as the number of themes in common divided by the number of different themes found by either MRM or pharmacists. For the 570 patients Jaccard coefficients ranged from a minimum of 0 to a maximum of 1, with a mean of 0.092 ± 0.117.

The top five themes by number of patients in common are shown in Table 3. Not surprisingly several of the most common themes found align with common health conditions in this cohort, namely hyperlipidemia and osteoporosis.

Some of the problems that can be identified by the software are shown in Tables 3 and 4. Table 3 shows there is some overlap of the ability of MRM to find the same kind of problems as pharmacists in the same patients. However, both pharmacists and MRM find many instances of the same problem in different patients. Table 4 shows examples of some of the themes at the extremes of overlap. The two example themes *calcium-channel blocker and reflux* and *anti-lipidemic drug, no indication* were identified in many patients by MRM but only once each by pharmacists. Similarly, the two example themes *vitamin, no indication* and *combine medications into combination product* illustrate that pharmacists identified many patients with particular problems that MRM could not identify.

Table 3. Top five themes by patients in common

Top five themes by cases in common	**Patients MRM found**	**Patients pharmacist found**	**Patients in common**	**Total Patients: pharmacists + MRM**
Osteoporosis (or risk) may require calcium and or vitamin D	137	117	49	205
Renal impairment and using (or check dose for) renally excreted drugs	122	48	24	146
Hyperlipidemia under/untreated	83	31	20	94
Sedatives long-acting or sedative long term	55	31	18	68
NSAID not recommended (heart disease/risk of bleed/other)	59	28	17	70

Table 4. Themes skewed in favour of MRM or pharmacists

Skewed themes with cases in common	Patients MRM found	Patients pharma-cist found	Patients in common	Total Patients: pharmacists + MRM
Calcium channel blocker and reflux	120	1	1	120
Anti-lipidemic drug, no indi-cation	56	1	1	56
Vitamin, no indication	1	6	1	6
Combine medications into combination product	3	10	1	12

5 Discussion

The majority of the unique pharmacist themes involved non-classifiable, mostly drug cost and compliance, problems. These pharmacist-only themes were not captured in the knowledge domain model. Although the majority of unique MRM themes could have been identified by pharmacists they were not. This was not due to lack of information on the part of pharmacists but more likely to be due to pharmacists having additional knowledge that rendered these issues moot. It is also possible that pharmacists were not aware of or simply missed these particular issues. Alternatively, the software may have produced erroneous findings.

The wide variety of variables including temporal context encapsulated in the model were manifested in the broad scope of problems that could be identified by the software. For 68 themes (out of 100 themes identified by MRM) the software showed the ability to identify the same issues that pharmacists could find in the same patients. In some circumstances half to all instances of a theme identified by pharmacists was also identified by MRM; most of the themes shown in Table 3 are examples of this.

The broad scope of themes and similarity of identification of themes in the same patients as pharmacists is encouraging; however, there were many patients who had particular problems identified by either MRM or pharmacists but not by both. Further, twenty-two themes were identified by MRM and by pharmacists without any patients in common. Several explanations are posited to account for these differences.

The first and main point is knowledge not captured and subsequently not able to be utilized by the software. Extending this point, knowledge may have been available but not entered into the software because it was not recorded anywhere by either the patient's GP or the reviewing pharmacist. Several themes stated some drugs had no indication for use because no suitable diagnosis was assigned to those patients. An example in Table 4, *anti-lipidemic drug, no indication*, shows MRM found many instances of this potential problem but pharmacists did not identify this as an issue. Does this mean pharmacists were aware of the indication for the drug? Or does it suggest pharmacists missed the opportunity to identify unnecessary medication?

Overall MRM found more problems than pharmacists. It is not unreasonable to suggest pharmacists may lack consistency in identifying DRPs. Correspondingly, it is not unreasonable to suggest MRM exemplifies consistency, as it is after all computer software. Several studies examining clinical decision support, including two prototypes on which MRM was based, have identified that humans lack consistency or lack the capacity to identify all relevant problems in contrast with the software [7, 8, 20]. Additionally, pharmacists may have focused on more important DRPs through prioritizing more pertinent DRP findings and ignoring lesser issues.

MRM did find substantially more problems than pharmacists, which raises some concerns about potential alert fatigue, a known limitation of many clinical decision support systems, wherein the system identifies so many irrelevant problems that the user simply ignores it entirely. It should be noted a portion of MRMs findings were duplications, 256 of 3209 DRPs. The central requirement and unfortunately concomitant problem of clinical decision support is the need to have sufficient information to present findings in context of the patient's current clinical situation. The application of MCRDR attempts to address the problem of context through incorporation of an extensive array of variables integrated with a knowledge base of many patient cases and inference rules.

However, it appears that MRM may not suffer from alert fatigue, as separate research that we have conducted, concerning the clinical relevance of the DRP findings of MRM and of pharmacists, was recently completed [11]. In that study experts in the field were of the opinion that both MRM and pharmacists identified clinically relevant DRPs [11]. That study supports the position that MRM may be more consistent than pharmacists by identifying a greater number of issues that pharmacists did not identify. Secondly, and importantly, despite the larger number of issues identified by MRM, lack of clinical relevance did not appear to be a factor.

A specific advantage of this implementation of MCRDR was the use of case-based reasoning, allowing the knowledge domain expert to readily add new rules and refine existing rules. This method incrementally increases the precision of rules in context of the uniquely varied situations encountered through amassing knowledge of individual patients. This is an important point, as the development of new medications, or new applications of existing medications, and ever expanding medical knowledge needs to be to be incorporated into such software on an ongoing basis to maintain the relevance of the knowledge base.

Due to the ability to easily add and refine the rules and knowledge-base a follow-up study may produce different, likely improved results. A subsequent investigation applying the same patient cases to the software and comparing the differences may be performed to determine whether DRP identification can be further enhanced over time.

MRM appears to work well in the HMR domain, but improvements may include a greater extent of variables such as compliance or cost-related concepts to widen problem detection scope as well as increasing accuracy of problem identification. Rule refinement to reduce the occurrence of duplicated DRPs is warranted. Another potential issue involves medication classification which was based on the ATC classification system. The ATC classification system included codes for combination products.

There may be limitations when attempting to create rules based on individual ingredients within combination products as each individual ingredient is not uniquely identified. Additionally, with the impending implementation of national electronic health record standards, data entry limitations such as transcription errors or missed data entry may be minimized by implementing these standards.

6 Conclusion

The use of ripple-down rules in this software did perform well in the complex and detailed HMR knowledge domain. It showed a reasonable degree of similarity with the human experts in the both the range of problem types that could be identified within its scope of knowledge, and in the frequency of problems found. MRM cannot find some of the problems that pharmacists could find, some things will always be missed because of incomplete data.

The truly interesting aspect is the software's capacity to identify more problems than pharmacists. This capacity to identify more problems did not appear to involve lack of relevance, but it is likely to be a strong indication of the consistent methodical ability of the machine to identify problems. This finding alone justifies the use of such a tool. MRM cannot replace pharmacists but may help pharmacists make good decisions and avoid missing important problems.

7 Competing interests

The author Gregory Peterson is an investor in Medscope Pty Ltd which developed MRM. The MRM software was based on the work of author Ivan Bindoff. Gregory Peterson was involved with the work of Ivan Bindoff as researcher and supervisor. Peter Tenni, a researcher previously involved with Ivan Bindoff's work, is currently the manager of the clinical division of Medscope Pty Ltd.

8 References

1. Pharmaceutical Care Network Europe, www.pcne.org/sig/drp/drug-related-problems.php
2. Home Medicines Review (HMR), www.medicareaustralia.gov.au/provider/pbs/fifth-agreement/home-medicines-review.jsp
3. Medicare Australia – Statistics – Item Reports, www.medicareaustralia.gov.au/statistics/mbs_item.shtml
4. Pharmaceutical Society of Australia, Guidelines for pharmacists providing home medicines review (HMR) services. Pharmaceutical Society of Australia (2011)
5. Basger, B.J., T.F. Chen, and R.J. Moles, Inappropriate medication use and prescribing indicators in elderly Australians: Development of a prescribing indicators tool. Drugs Aging. 25(9), 777-793 (2008)
6. Medscope Medication Review Mentor (MRM), www.medscope.com.au

7. Bindoff, I., Stafford, A., Peterson, G., Kang, B.H., Tenni, P.: The potential for intelligent decision support systems to improve the quality and consistency of medication reviews. J Clin Pharm Ther. 37(4), 452-458 (2011)
8. Bindoff, I.K., Tenni, P.C., Peterson, G.M., Kang, B.H., Jackson, S.L.: Development of an intelligent decision support system for medication review. J Clin Pharm Ther. 32(1), 81-88 (2007)
9. Compton, P., Peters, L., Edwards, G., Lavers, T.G.: Experience with Ripple-Down Rules. Knowledge-Based Systems. 19(5), 356-362 (2006)
10. Curtain, C., Westbury, J., Bindoff, I., Peterson, G.: Validation of home medicines review decision support software. In Graduate research - Sharing excellence in research conference proceedings, p. 23 Hobart (2012)
11. Curtain, C., Bindoff, I., Westbury, J., Peterson, G.: Validation of decision support software for identification of drug-related problems. In 11th National conference of Emerging Researchers in Ageing, In Press, Brisbane (2012)
12. Curtain, C., Bindoff, I., Westbury, J., Peterson, G.: Can software assist the home medicines review process by identifying clinically relevant drug-related problems? In ASCEPT-APSA 2012 conference. In Press. Sydney (2012)
13. Stafford, A., Tenni, P., Peterson, G., Doran, C., Kelly, W.: IIG-021 - VALMER (the Economic Value of Home Medicines Reviews), Pharmacy Guild of Australia
14. WHO Collaborating Centre for Drug Statistics Methodology Norwegian Institute of Public Health. International language for drug utilization research ATC / DDD, www.whocc.no
15. Jamoulle, M. ICPC2, the international classification of primary care, www.ulb.ac.be/esp/wicc/icpc2.html#C2
16. Tenni, P.: Manager, Clinical Division, Medscope, Hobart (2012)
17. Fick, D.M., Cooper, J.W., Wade, W.E., Waller, J.L., Maclean, J.R., Beers, M.H.: Updating the Beers criteria for potentially inappropriate medication use in older adults: results of a US consensus panel of experts. Arch Intern Med. 163, 2716-2724 (2003)
18. Gallagher, P., Ryan, C., Byrne, S., Kennedy, J., O'Mahony, D.: STOPP (Screening Tool of Older Person's Prescriptions) and START (Screening Tool to Alert doctors to Right Treatment). Consensus validation. Int J Clin Pharmacol Ther. 46(2), 72-83 (2008)
19. Williams, M., Peterson, G.M., Tenni, P.C., Bindoff, I.K., Stafford, A.C.: DOCUMENT: a system for classifying drug-related problems in community pharmacy. Int J Clin Pharm. 34(1), 43-52 (2011)
20. Martins, S.B., Lai, S., Tu, S., Shankar, R., Hastings, S.N., Hoffman, B.B., Dipilla, N., Goldstein, M.K.: Offline testing of the ATHENA Hypertension decision support system knowledge base to improve the accuracy of recommendations. AMIA Annu Symp Proc, 539-43 (2006)

FS-XCS vs. GRD-XCS: An analysis using high-dimensional DNA microarray gene expression data sets

Mani Abedini[1], Michael Kirley[1], and Raymond Chiong[1,2]

[1] Department of Computing and Information Systems,
The University of Melbourne, Victoria 3010, Australia
{mabedini,mkirley,rchiong}@csse.unimelb.edu.au
[2] Faculty of Higher Education Lilydale,
Swinburne University of Technology, Victoria 3140, Australia
rchiong@swin.edu.au

Abstract. XCS, a Genetic Based Machine Learning model that combines reinforcement learning with evolutionary algorithms to evolve a population of classifiers in the form of condition-action rules, has been used successfully for many classification tasks. However, like many other machine learning algorithms, XCS becomes less effective when it is applied to high-dimensional data sets. In this paper, we present an analysis of two XCS extensions – FS-XCS and GRD-XCS – in an attempt to overcome the dimensionality issue. FS-XCS is a standard combination of a feature selection method and XCS. As for GRD-XCS, we use feature quality information to bias the evolutionary operators without removing any features from the data sets. Comprehensive numerical simulation experiments show that both approaches can effectively enhance the learning performance of XCS. While GRD-XCS has obtained significantly more accurate classification results than FS-XCS, the latter has produced much quicker execution time than the former.

1 Introduction

Classification tasks arise in many areas of science and engineering. One such example is disease classification based on gene expression profiles in bioinformatics. Gene expression profiles provide important insights into, and further our understanding of, biological processes. They are key tools used in medical diagnosis, treatment, and drug design [21]. From a clinical perspective, the classification of gene expression data is an important problem and a very active research area (see [3] for a review). DNA microarray technology has advanced a great deal in recent years. It is possible to simultaneously measure the expression levels of thousands of genes under particular experimental environments and conditions [22]. However, the number of samples tends to be much smaller than the number of genes (features)[1]. Consequently, the high dimensionality of a given

[1] Generally speaking, the number of samples must be larger than the number of features for good classification performance.

data set poses many statistical and analytical challenges, which often degrade the performance of classification methods used.

XCS – the eXtended Classifier System – is a Genetic Based Machine Learning (GBML) method that has been successfully used for a wide variety of classification applications, including medical data mining. XCS can learn from sample data in multiple iterative cycles. This is a great characteristic, but it also exhibits two common pitfalls that most classification methods have: sensitivity to data noise and "the curse of dimensionality" [22]. Both issues can easily jeopardise the learning process. A well-known solution is to use a cleansing stage. For example, feature selection/ranking techniques can remove unnecessary features from the data set. Reducing the dimensionality and removing noisy features can improve learning performance. Nevertheless, there exist data sets with highly co-expressed features, such as those studying Epistasis phenomena, that do not allow effective feature reduction. Examples of this include protein structure prediction and protein-protein interaction.

In this paper, we study two extensions of XCS inspired by feature selection techniques commonly used in machine learning: FS-XCS with effective feature reduction in place and GRD-XCS [1] that does not remove any features. The proposed model uses some prior knowledge, provided by a feature ranking method, to bias the discovery operators of XCS. A series of comprehensive numerical experiments on high-dimensional medical data sets has been conducted. The results of these simulation experiments suggest that both extensions can effectively enhance the XCS's learning performance. While GRD-XCS has performed significantly more accurate than FS-XCS, the latter is shown to have much quicker execution time compared to the former.

The remainder of this paper is organised as follows: Section 2 briefly describes some related work on XCS. In Section 3, we present the details of our proposed model. Section 4 discusses the experimental settings and results. Finally, we draw conclusion and highlight future possibilities in Section 5.

2 Related Work

GBML concerns applying evolutionary algorithms (EAs) to machine learning. EAs belong to the family of nature-inspired optimisation algorithms [9, 10]. As a manifestation of population-based, stochastic search algorithms that mimic natural evolution, EAs use genetic operators such as crossover and mutation for the search process to generate new solutions through a repeated application of variation and selection [11].

It is well documented in the evolutionary computation literature that the implementation of EA's genetic operators can influence the trajectory of the evolving population. However, there has been a paucity of studies focused specifically on the impact of selected evolutionary operator implementations in Learning Classifier Systems (LCSs), a type of GBML algorithm for rule induction. Here, we briefly describe some of the key studies related to LCSs in general and XCS – a Michigan-style LCS – in particular.

In one of the first studies focused on the rule discovery component specifically for XCS, Butz et al. [7] have shown that uniform crossover can ensure successful learning in many tasks. In subsequent work, Butz et al. [6] introduced an informed crossover operator, which extended the usual uniform operator such that exchanges of effective building blocks occurred. This approach helped to avoid the over-generalisation phenomena inherent in XCS [14]. In other work, Bacardit et al. [4] customised the GAssist crossover operator to switch between the standard crossover or a new simple crossover, SX. The SX operator uses a heuristic selection approach to take a minimum number of rules from the parents (more than two), which can obtain maximum accuracy. Morales-Ortigosa et al. [16] have also proposed a new XCS crossover operator, BLX, which allowed for the creation of multiple offspring with a diversity parameter to control differences between offspring and parents. In a more comprehensive overview paper, Morales-Ortigosa et al. [17] presented a systematic experimental analysis of the rule discovery component in LCSs. Subsequently, they developed crossover operators to enhance the discovery component based on evolution strategies with significant performance improvements.

Other work focused on biased evolutionary operators in LCSs include the work of Jos-Revuelta [18], who introduced a hybridised Genetic Algorithm-Tabu Search (GA-TS) method that employed modified mutation and crossover operators. Here, the operator probabilities were tuned by analysing all the fitness values of individuals during the evolution process. Wang et al. [20] used *Information Gain* as part of the fitness function in an EA. They reported improved results when comparing their model to other machine learning algorithms. Recently, Huerta et al. [5] combined *linear discriminant analysis* with a GA to evaluate the fitness of individuals and associated discriminate coefficients for crossover and mutation operators. Moore et al. [15] argued that biasing the initial population, based on expert knowledge preprocessing, would lead to improved performance of the evolutionary based model. In their approach, a statistical method, *Tuned ReliefF*, was used to determine the dependencies between features to seed the initial population. A modified fitness function and a new guided mutation operator based on features dependency was also introduced, leading to significantly improved performance.

3 The Model

We have designed and developed two extensions of XCS, both inspired by feature selection techniques commonly used in machine learning. The first extension, which we call FS-XCS, is a combination of a Feature Selection method and the original XCS. The second extension, which we call GRD-XCS, incorporates a probabilistically Guided Rule Discovery mechanism for FS-XCS. The motivation behind both extensions was to improve classifier performance (in terms of accuracy and execution time), especially for high-dimensional classification problems.

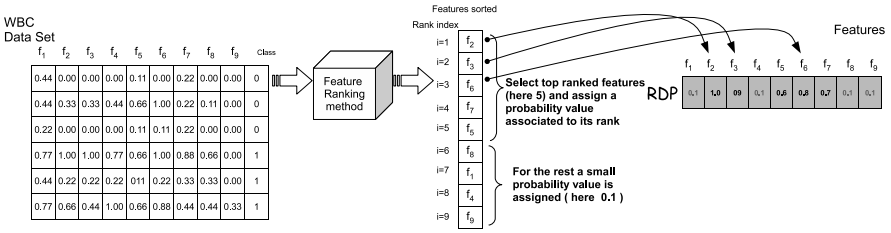


Fig. 1. Here, Information Gain has been used to rank the features. The top Ω features (in this example $\Omega = 5$) are selected and allocated relatively large probability values $\in [\gamma, 1]$. The RDP vector maintains these values. The probability value of the highest ranked feature is set to 1.0. Other features receive smaller probability values relative to their rank (in this example γ =0.5). Features that are not selected based on Information Gain are assigned a very small probability value (in this example $\xi = 0.1$).

FS-XCS uses feature ranking methods to reduce the dimension of a given data set before XCS starts to process the data set. It is a fairly straightforward hybrid approach. However, in GRD-XCS information gathered from feature ranking methods is used to build a probability model that biases the evolutionary operators of XCS. The feature ranking probability distribution values are recorded in a Rule Discovery Probability (RDP) vector. Each value of the RDP vector ($\in [0, 1.0]$) is associated with a corresponding feature. The RDP vector is then used to bias the feature-wise uniform *crossover*, *mutation*, and *don't care* operators, which are part of the XCS rule discovery component.

The actual values in the RDP vector are calculated based on the rank of the corresponding feature as described below:

$$RDP_i = \begin{cases} \frac{1-\gamma}{\Omega} \times (\Omega - i) + \gamma & \text{if } i \leq \Omega \\ \xi & \text{otherwise} \end{cases} \quad (1)$$

where i represents the rank index in ascending order for the selected top ranked features Ω. The probability values associated with the top ranked features would be some relatively large values ($\in [\gamma, 1]$) depending on the feature rank; for the others a very low probability value ξ is given. Thus, all features have a chance to participate in the rule discovery process. However, the Ω-top ranked features have a greater chance of being selected (see Figure 1 for an example).

GRD-XCS uses the probability values recorded in the RDP vector in the preprocessing phase to bias the evolutionary operators used in the rule discovery phase of XCS. The modified algorithms describing the *crossover*, *mutation* and *don't care* operators in GRD-XCS are very similar to standard XCS operators:

- GRD-XCS *crossover* operator: This is a hybrid uniform/n-point function. An additional check of each feature is carried out before the exchange of genetic material. If $Random[0, 1) < RDP[i]$ then feature i is swapped between the selected parents (Algorithm 1).

Algorithm 1 Guided Uniform Crossover algorithm

Require: Individuals: Cl_1,Cl_2 $\in[A]$, Probability Vector: RDP, Crossover Probability: χ

```
if Random[0,1) < χ then
  for i = 1 To SizeOf(Features) do
    if Random[0,1) < RDP[i] then
      SWAP(Cl1[i],Cl2[i])
    end if
  end for
end if
```

Algorithm 2 Guided Mutation algorithm

Require: Individual: Cl $\in[A]$, Probability Vector: RDP, Mutation Probability: μ

```
for i = 1 To SizeOf(Features) do
  if Random[0,1) < RDP[i] ×μ then
    Mutate(Cl[i])
  end if
end for
```

Algorithm 3 Guided Don't Care algorithm

Require: Individuals: Cl $\in[A]$, Probability Vector: RDP, Don't Care Probability: $P_{\#}$

```
for i = 1 To SizeOf(Features) do
  if Random[0,1) < (1 − RDP[i]) ×P# then
    P1[i] ← #
  end if
end for
```

- GRD-XCS *mutation* operator: It uses the RDP vector to determine if feature i is to undergo mutation; the base-line mutation probability is multiplied by RDP for each feature. Therefore, the mutation probability is not a uniform distribution anymore. The more informative features have better chance to be selected for mutation (Algorithm 2).
- GRD-XCS *don't care* operator: In this special mutation operator, the values in the RDP vector are used in the reverse order. That is, if feature i has been selected to be mutated and $Random[0,1) < (1 - RDP[i])$, then feature i is changed to # ("don't care") (see Algorithm 3).

The application of the RDP vector reduces the crossover and mutation probabilities for "uninformative" features. However, it increases the "don't care" operator probability for the same feature. Therefore, the more informative features should appear in rules more often than the "uninformative" ones.

4 Experiments and Results

We have conducted a series of independent experiments to compare the performance of FS-XCS and GRD-XCS. A suite of feature selection techniques have

Table 1. Data set details

Data Set	#Instances	#Features	Cross Validation	Reference
High-dimensional data sets (Microarray DNA gene expression)				
Breast cancer	22	3226	3	[13]
Colon cancer	62	2000	10	[2]
Leukemia cancer	72	7129	10	[12]
Prostate cancer	136	12600	10	[19]

been tested: Correlation based Feature Selection (CFS), Gain Ratio, Information Gain, One Rule, ReliefF and Support Vector Machine (SVM). Four DNA microarray gene expression data sets have been used in the experiments. The details of these data sets are reported in Table 1.

Our algorithms were implemented in C++, based on the Butz's XCS code[2]. The WEKA package (version 3.6.1)[3] was used for feature ranking. All experiments were performed on the *VPAC*[4] Tango Cluster server. Tango has 111 computing nodes. Each node is equipped with two 2.3 GHz AMD based quad core Opteron processors, 32GB of RAM and four 320GB hard drives. Tango's operating system is the Linux distribution CentOS (version 5).

4.1 Parameter settings

Default parameter values as recommended in [8] have largely been used to configure the underlying XCS model. For parameters specific to our proposed model, we have carried out a detailed analysis to determine the optimal settings. In particular, we have tested a range of Ω values $\Omega = 10$, 20, 32, 64, 128, 256 and population sizes $pop\_size = 500$, 1000, 2000, 5000. The analysis suggested that $\Omega = 20$ with a population size of 2000 can provide an acceptable accuracy level within reasonable execution time for FS-XCS. As for GRD-XCS, the setting of $\Omega = 128$ and $pop\_size = 500$ was found to have produced the best results. As such, these parameter values were used for the results presented in Section 4.3.

The limits used in probability value calculations in Equation 1 were set to $\gamma = 0.5$ and $\xi = 0.1$. In all experiments, the number of iterations was capped at 5000.

4.2 Evaluation

For each scenario (parameter value–data set combination), we performed N-fold cross validation experiments over 100 trials (see Table 1). The average accuracy

[2] The source code is available at the Illinois Genetic Algorithms Laboratory (IlliGAL) site http://www.illigal.org/

[3] Weka 3 is an open source data mining tool (in Java), with a collection of machine learning algorithms developed by the Machine Learning Group at University of Waikato – http://www.cs.waikato.ac.nz/ml/weka/

[4] Victorian Partnership for Advanced Computing: www.vpac.org

Table 2. Average accuracy (measured by AUC values) of the base-line XCS, FS-XCS and GRD-XCS on all selected microarray gene expression data sets.

base-line XCS	FS-XCS	GRD-XCS
0.77	0.88	0.98

values for specific parameter combinations have been reported using the Area Under the ROC Curve – the AUC value. The ROC curve is a graphical way to depict the tradeoff between the *True Positive rate* (TPR) on the Y axis and the *False Positive rate* (FPR) on the X axis. The AUC values obtained from the ROC graphs allow for easy comparison between two or more plots. Larger AUC values represent higher overall accuracy.

Appropriate statistical analyses using paired t-tests were conducted to determine whether there were statistically significant differences between particular scenarios in terms of both accuracy and execution time. Scatter plots of the observed and fitted values and Q-Q plots were used to verify normality assumptions. These statistical analyses were performed using the IBM SPSS Statistics (version 19) software.

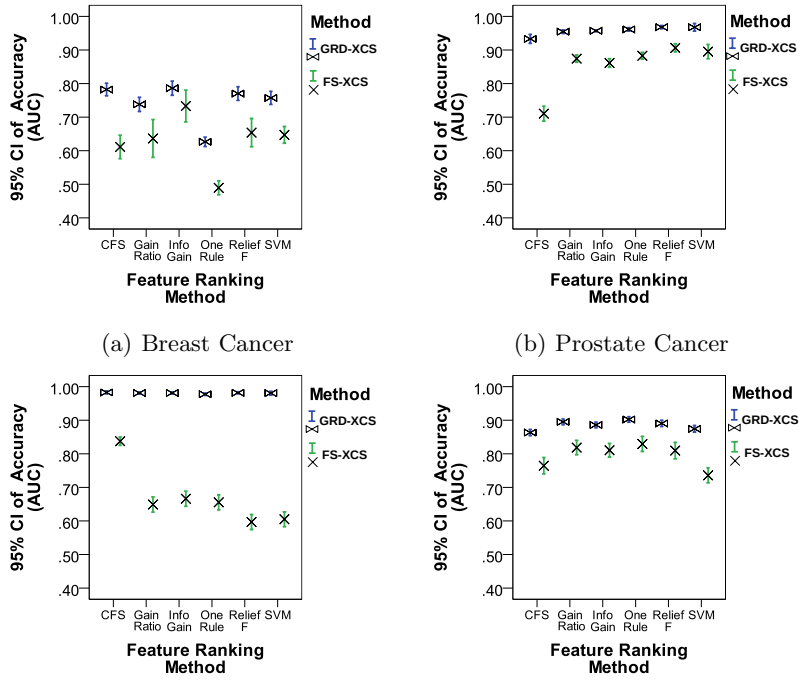
4.3 FS-XCS vs. GRD-XCS

To begin with, we have compared the average accuracy of FS-XCS and GRD-XCS with the base-line XCS (without feature selection) using all the aforementioned feature ranking methods on the microarray gene expression data sets listed in Table 1. The results, as shown in Table 2, indicate that GRD-XCS has an overall better accuracy than FS-XCS: the average FS-XCS accuracy using various feature selection techniques is 0.88 while the average accuracy of GRD-XCS using the same feature ranking methods is 0.98. Meanwhile, both FS-XCS and GRD-XCS are better than the base-line XCS – the latter has managed only an average accuracy of 0.77. For the rest of this section, we will focus on a detailed comparison between FS-XCS and GRD-XCS.

Figure 2 shows the AUC values of FS-XCS and GRD-XCS when different feature ranking methods are used. From the figure, it is clear that GRD-XCS is significantly more accurate than FS-XCS. The accuracy result of both FS-XCS and GRD-XCS for every feature ranking method, except Information Gain over the Breast Cancer data set, is significantly different ($p < 0.001$).

In Figure 3, FS-XCS is shown to be significantly faster than GRD-XCS ($p < 0.001$) in terms of execution time. This is much expected since FS-XCS works with only a fraction of the original data set size (i.e., 20 features) while GRD-XCS still accepts the entire data set with thousands of features. The only exception is when Gain Ratio has been applied over the Breast Cancer data set – in this case there is strong evidence that both FS-XCS and GRD-XCS have significantly equal average execution time ($p = 0.94$).

Figures 4 and 5 depict some general insight into the population diversity. In the majority of cases, GRD-XCS has less diversity.



(a) Breast Cancer

(b) Prostate Cancer

(c) Leukemia

(d) Colon Cancer

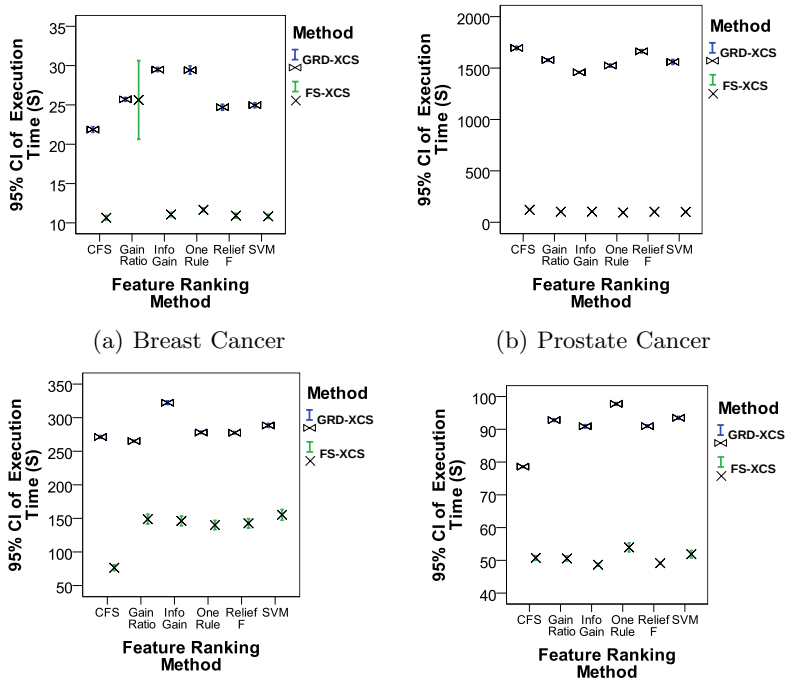
Fig. 2. The accuracy (AUC) of FS-XCS vs. GRD-XCS when various feature ranking methods are applied.

The average length of each classifier in GRD-XCS is almost always significantly smaller than FS-XCS ($p < 0.05$). The significant similar cases are Gain Ratio ($p = 0.80$) and ReliefF ($p = 0.26$) on the Prostate Cancer data set.

The average number of macro classifiers in GRD-XCS is significantly smaller than the average number of macro classifiers in FS-XCS. As can be seen in Figures 5(b) and (d), the difference is getting more obvious when the dimensionality increases (for Prostate Cancer and Colon Cancer). However, there is a different story for the Breast Cancer data set where the average number of macro classifiers in the GRD-XCS population is larger than FS-XCS. It would be a fair conclusion to say that GRD-XCS is exploring the solution space in a more focused manner than FS-XCS. In other words, the guided rule discovery approach forces the learning process to generate less diverse testing hypothesis; however this behaviour can evolve more accurate classifiers.

5 Conclusion and Future Work

In this paper, we have analysed the performance of FS-XCS and GRD-XCS based on some high-dimensional classification problems. Comprehensive numer-



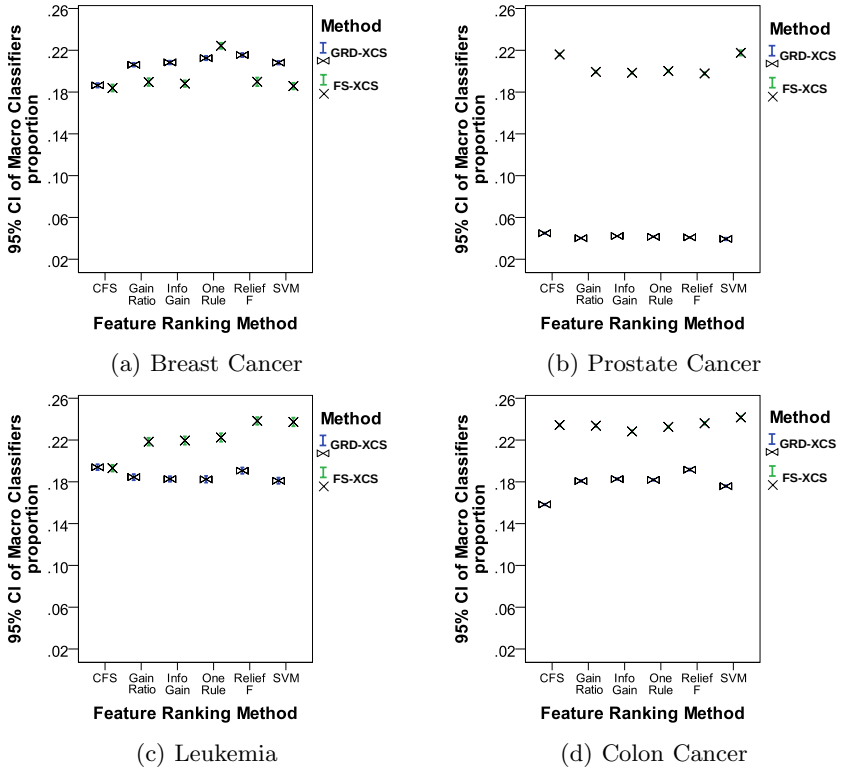
(a) Breast Cancer (b) Prostate Cancer

(c) Leukemia (d) Colon Cancer

Fig. 3. The execution time (in seconds) of FS-XCS vs. GRD-XCS when various feature ranking methods are applied.

ical simulations have established that GRD-XCS is significantly more accurate than FS-XCS in terms of classification results. On the other hand, FS-XCS is significantly faster than GRD-XCS in terms of execution time. The results of FS-XCS suggest that normally 20 top-ranked features would be enough to build a good classifier, although this classifier is significantly less accurate than the equivalent GRD-XCS model. Nevertheless, both models have performed better than the base-line XCS.

To sum up, using feature selection to highlight the more informative features and using them to guide the XCS rule discovery process is better than applying feature reduction approaches. This is mainly due to the fact that GRD-XCS can transform poor classifiers (created from the uninformative features) into highly accurate classifiers. From the empirical analysis presented it is clear that the performance of different feature selection techniques varies inevitably depending on the data set characteristic. Future work will therefore attempt to rectify this through the idea of ensemble learning. That is, we can build an ensemble classifier from multiple XCS based models (may it be FS-XCS or GRD-XCS). Each of these XCS cores can use a distinctive feature selection method. The



(a) Breast Cancer (b) Prostate Cancer

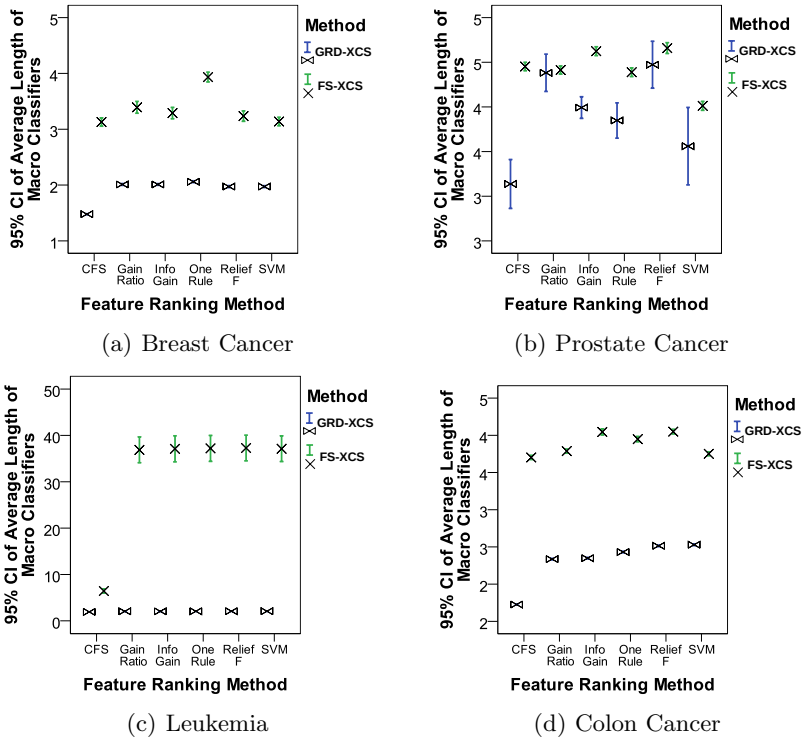
(c) Leukemia (d) Colon Cancer

Fig. 4. The proportion of macro classifiers to the population size of FS-XCS vs. GRD-XCS when various feature ranking methods are applied.

results of all XCS cores are then combined to form the ensemble result – for instance by using a majority voting technique.

References

1. M. Abedini and M. Kirley. An enhanced XCS rule discovery module using feature ranking. *International Journal of Machine Learning and Cybernetics*, 10.1007/s13042-012-0085-9, 2012.
2. U. Alon, N. Barkai, D. A. Notterman, K. Gishdagger, S. Ybarradagger, D. Mackdagger, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. of the National Academy of Sciences of the USA*, 96:6745–6750, 1999.
3. M. H. Asyali, D. Colak, O. Demirkaya, and M. S. Inan. Gene expression profile classification: A review. *Current Bioinformatics*, 1(1):55–73, 2006.
4. J. Bacardit and N. Krasnogor. Smart crossover operator with multiple parents for a Pittsburgh learning classifier system. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, pages 1441–1448. ACM Press, 2006.



(a) Breast Cancer (b) Prostate Cancer

(c) Leukemia (d) Colon Cancer

Fig. 5. The average length of macro classifiers (rules) of FS-XCS vs. GRD-XCS when various feature ranking methods are applied.

5. E. Bonilla Huerta, J. C. Hernandez Hernandez, and L. A. Hernandez Montiel. A new combined filter-wrapper framework for gene subset selection with specialized genetic operators. In *Advances in Pattern Recognition*, volume 6256 of *Lecture Notes in Computer Science*, pages 250–259. Springer, 2010.
6. M. Butz, M. Pelikan, X. Lloral, and David E. Goldberg. Automated global structure extraction for effective local building block processing in XCS. *Evolutionary Computation*, 14(3):345–380, 2006.
7. M. V. Butz, D. E. Goldberg, and K. Tharakunnel. Analysis and improvement of fitness exploitation in XCS: Bounding models, tournament selection, and bilateral accuracy. *Evolutionary Computation*, 11(3):239–277, 2003.
8. M. V. Butz and S. W. Wilson. An algorithmic description of XCS. *Soft Computing*, 6(3–4):144–153, 2002.
9. R. Chiong, editor. *Nature-Inspired Algorithms for Optimisation*. Springer, 2009.
10. R. Chiong, F. Neri, and R. I. McKay. Nature that breeds solutions. In R. Chiong, editor, *Nature-Inspired Informatics for Intelligent Applications and Knowledge Discovery: Implications in Business, Science and Engineering*, chapter 1, pages 1–24. Information Science Reference, Hershey, PA, 2009.
11. R. Chiong, T. Weise, and Z. Michalewicz, editors. *Variants of Evolutionary Algorithms for Real-World Applications*. Springer, 2012.

12. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.
13. I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, P. Meltzer, B. Gusterson, M. Esteller, O. P. Kallioniemi, B. Wilfond, A. Borg, and J. Trent. Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medicine*, 344(8):539–548, 2001.
14. P. L. Lanzi. A study of the generalization capabilities of XCS. In Thomas Bäck, editor, *Proceedings of the 7th International Conference on Genetic Algorithms*, pages 418–425. Morgan Kaufmann, 1997.
15. J. H. Moore and B. C. White. Exploiting expert knowledge in genetic programming for genome-wide genetic analysis. In *PPSN*, volume 4193 of *Lecture Notes in Computer Science*, pages 969–977. Springer, 2006.
16. S. Morales-Ortigosa, A. Orriols-Puig, and E. Bernadó-Mansilla. New crossover operator for evolutionary rule discovery in XCS. In *Proceedings of the 8th International Conference on Hybrid Intelligent Systems*, pages 867–872. IEEE Computer Society, 2008.
17. S. Morales-Ortigosa, A. Orriols-Puig, and E. Bernadó-Mansilla. Analysis and improvement of the genetic discovery component of XCS. *International Journal of Hybrid Intelligent Systems*, 6(2):81–95, 2009.
18. L. M. San Jose-Revuelta. *A Hybrid GA-TS Technique with Dynamic Operators and its Application to Channel Equalization and Fiber Tracking*. I-Tech Education and Publishing, 2008.
19. D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, and A. A. Renshaw. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.
20. P. Wang, T. Weise, and R. Chiong. Novel evolutionary algorithms for supervised classification problems: An experimental study. *Evolutionary Intelligence*, 4(1):3–16, 2011.
21. F.-X. Wu, W. J. Zhang, and A. J. Kusalik. On determination of minimum sample size for discovery of temporal gene expression patterns. In *Proceedings of the First International Multi-Symposiums on Computer and Computational Sciences*, pages 96–103, 2006.
22. Y. Zhang and J. C. Rajapakse. *Machine Learning in Bioinformatics*. Wiley Series in Bioinformatics. 1'st edition, 2008.

Reliable Epileptic Seizure Detection Using an Improved Wavelet Neural Network

Zarita Zainuddin[1,*], Lai Kee Huong[1], and Ong Pauline[1]

[1]School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia.
zarita@cs.usm.my, laikeehuong1986@yahoo.com,
ong.pauline@hotmail.com

Abstract. Electroencephalogram (EEG) signals analysis is indispensable in epilepsy diagnosis as it offers valuable insights for locating the abnormal distortions in the brain wave. However, visual interpretation of the massive amount of EEG signals is time-consuming, and there is often inconsistent judgment between the experts. Thus, a reliable seizure detection system is highly sought after. A novel approach for epileptic seizure detection is proposed in this paper, where the statistical features extracted from the discrete wavelet transform are used in conjunction with an improved wavelet neural network in order to identify the occurrence of seizures. Experimental simulations were carried out on a well-known publicly available dataset, which was kindly provided by Ralph Andrzejak from the Epilepsy center in Bonn, Germany. The obtained high prediction accuracy, sensitivity and specificity demonstrated the feasibility of the proposed seizure detection scheme.

Keywords: Epileptic seizure detection, fuzzy *C*-means clustering, *K*-means clustering, type-2 fuzzy *C*-means clustering, wavelet neural networks.

1 Introduction

Since its first inception reported by German neuropsychiatrist Hans Berger in the year 1924, the electroencephalogram (EEG) signals, which record the electrical activity in the brain, have emerged as an essential alternative in diagnosing neurological disorders. By analyzing the EEG recordings, inherent information from different physiological states of the brain can be extracted, which are extremely crucial for the epileptic seizure detection since the occurrence of seizure exhibits clear transient abnormalities in the EEG signals. Thus, a warning signal can be initiated in time to avoid any unwanted seizure related accidents and injuries, upon detecting an impending seizure attack.

While vital as a ubiquitous tool which supports general diagnostic of epilepsy, the clinical implementation of EEG is constrained due to the challenges of: (i) Available therapies require long term continuous monitoring of EEG signals. The generated massive amounts of EEG recordings have to be painstakingly scanned and analyzed visually by neurophysiologists, which is a tedious and time-consuming task. (ii) There

often is disagreement among different physicians during the analysis of ictal signals [19]. Undoubtedly, an automated diagnostic system that is capable of distinguishing the transient patterns of epileptiform activity from the EEG signals with reliable precision is of great significance.

Various efforts have been devoted in the literature in this regard. Generally speaking, a typical epileptic seizure detection process consists of two stages wherein, the inherent information that characterizes the different states of brain electrical activity are first derived from the EEG recordings using some feature extraction techniques, and subsequently, a chosen expert system is trained based on the obtained features. The discrete wavelet transform (DWT) has gained practical interest in extracting the valuable information embedded on the EEG signals due to its ability in capturing precise frequency information at low frequency bands and time information at high frequency bands [4], [9], [22], [25]. EEG signals are non-stationary in nature, and they contain high frequency information with short time period and low frequency information with long time period [18]. Therefore, by analyzing the biomedical signals at different time and frequency resolutions, DWT is able to preprocess the biomedical signals efficiently in the feature extraction stage.

In the second stage of the seizure detection scheme, a great deal of different artificial neural networks (ANNs) based expert systems have been utilized extensively in the emerging field of epilepsy diagnosis. For instance, the multilayer perceptrons, radial basis function neural networks, support vector machines, probabilistic neural networks, and recurrent neural networks are some of the models that have been previously reported in literature [5], [12], [14], [19], [22]. ANNs are powerful mathematical models that are inspired from their biological counterparts - the biological neural networks, which concern on how the interconnecting neurons process a massive amount of information at any given time. The utilization of ANNs in the seizure detection study is appropriate in nature, due to their capability of finding the underlying relationship between rapid variations in the EEG recordings, in addition to having the characteristics of fault tolerance, massive parallel processing ability, and adaptive learning capability.

The objective of this paper is to present a novel scheme based on an improved WNNs for the optimal classification of epileptic seizures in EEG recordings. The normal as well as the epileptic EEG signals were first pre-processed using the DWT wherein, the signals were decomposed into several frequency subbands. Subsequently, a set of statistical features were extracted from each frequency subband, and was used as a feature set to train a wavelet neural networks (WNNs) based classifier. It is worth mentioning that the feature selection of EEG signals using DWT and epileptic seizure detection with ANNs are well-accepted methodologies by medical experts [6-7].

The paper is organized as follows. In Section 2, the clinical data used in this study is first presented, followed by the feature extraction method based on the DWT. The implementation of the improved WNNs is next described in Section 3. In Section 4, the effectiveness of the proposed WNNs in epileptic seizure detection is presented and finally, conclusions are drawn in Section 5.

2 Materials and Methods

The flow of the methodology used in this study is depicted in the block diagram in Fig. 1, which will be discussed in detail in the following sections.

2.1 Clinical data selection

The EEG signals used in this study were acquired from a publicly available benchmark dataset [2]. The dataset is divided into five sets, labeled set A until E. Each set of the data consists of 100 segments, with each segment being a time series with 4097 data points. Each segment was recorded for 23.6 s at a sampling rate of 173.61 Hz. Each of the five sets was recorded under different circumstances. Both sets A and B were recorded from healthy subjects, with set A recorded with their eyes open whereas set B with their eyes closed. On the other hand, sets C until E were obtained from epileptic patients. Set C and D were recorded during seizure free period, where set C was recorded from the hippocampal formation of the opposite hemisphere of the brain, whereas set D was obtained from within the epileptogenic zone. The last data set, set E, contains ictal data that were recorded when the patients were experiencing seizure. In other words, the first four sets of data, sets A until D, are normal EEG signals, while set E represents epileptic EEG signals.

2.2 Discrete wavelet transform for feature extraction

DWT offers a more flexible time-frequency window function, which narrows when observing high frequency information and widens when analyzing low frequency resolution. It is implemented by decomposing the signal into coarse approximation and detail information by using successive low-pass and high-pass filtering, which is illustrated in Fig. 2.

As shown in this figure, a sample signal $x(n)$, is passed through the low-pass filter G_0 and high-pass filter H_0 simultaneously until the desired level of decomposition is reached. The low-pass filter produces coarse approximation coefficients $a(n)$, whereas the high-pass filter outputs the detail coefficients $d(n)$. The size of the approximation coefficients and detail coefficients decreases by a factor of 2 at each successive decomposition.

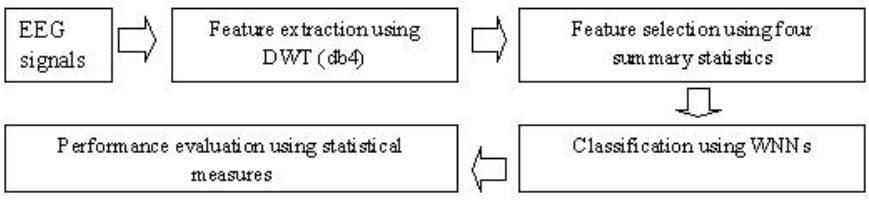


Fig. 1. Block diagram for the proposed seizure detection scheme.

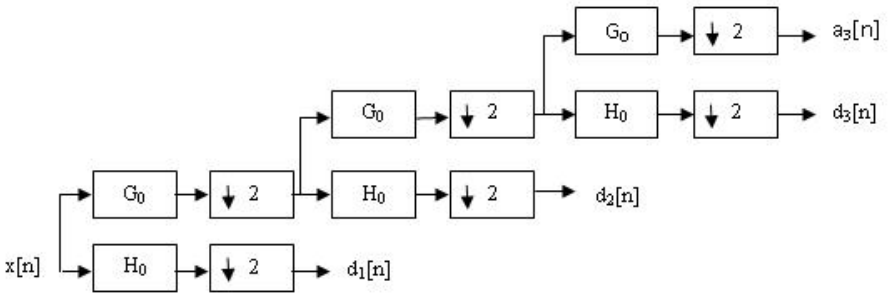


Fig. 2. A three-level wavelet decomposition tree.

Selecting the appropriate number of decomposition level is important for DWT. For the EEG signal analysis, the number of decomposition levels can be determined directly, based on their dominant frequency components. The number of levels is chosen in such a way that those parts of the signals which correlate well with the frequencies required for the classification of EEG signals are retained in the wavelet coefficients [17]. Since the clinical data used were sampled at 173.61Hz, the DWT using Daubechies wavelet of order 4 (db4), with four decomposition levels is chosen, as suggested in [21]. The db4 is suitable to be used as wavelets of lower order are too coarse to represent the EEG signals, while wavelets of higher order oscillate too wildly [1]. The four-level wavelet decomposition process will yield a total of five groups of wavelet coefficients, each corresponds to their respective frequency. They are d_1(43.4-86.8Hz), d_2(21.7-43.4Hz), d_3(10.8-21.7Hz), d_4(5.4-10.8Hz), and a_4(0-5.4Hz), which correlate with the EEG spectrum that fall within four frequency bands of: delta (1-4Hz), theta (4-8Hz), alpha (8-13Hz) and beta (13-22Hz).

Subsequently, the statistical features of these decomposition coefficients are extracted, which are:

1. The 90th percentile of the absolute values of the wavelet coefficients
2. The 10th percentile of the absolute values of the wavelet coefficients
3. The mean of the absolute values of the wavelet coefficients
4. The standard deviation of the wavelet coefficients.

It is worth mentioning that instead of the usual extrema (maximum and minimum of the wavelet coefficient), the percentiles are selected in this case in order to eliminate the possible outliers [11]. At the end of the feature extraction stage, a feature vector of length 20 is formed for each EEG signal.

3 Classification using an improved wavelet neural networks

WNNs are feedforward neural networks with three layers – the input layer, the hidden layer, and the output layer [26] . As the name suggests, the input layer receives input values and transmits them to the single hidden layer. The hidden nodes consist of

continuous wavelet functions, such as Gaussian wavelet, Mexican Hat wavelet, or Morlet wavelet, which perform the nonlinear mapping. The product from this hidden layer will then be sent to the final output layer.

Mathematically, a typical WNN is modeled by the following equation:

$$y(\mathbf{x}) = \sum_{i=1}^{p} w_{ij}\psi\left(\frac{\mathbf{x}-\mathbf{t}_i}{d}\right) + \mathbf{b}, \tag{1}$$

where y is the desired output, $\mathbf{x} \in \mathbb{R}^m$ is the input vector, p is the number of hidden neurons, w_{ij} is the weight matrix whose values will be adjusted iteratively during the training phase in order to minimize the error goal, ψ is the wavelet activation function, $\mathbf{t}$ is the translation vector, d is the dilation parameter, and $\mathbf{b}$ is the column matrix that contains the bias terms. The network structure is illustrated in Fig. 3.

The WNNs are distinct from those of other ANNs in the sense that [26]:

- WNNs show relatively faster learning speed owing to the constitution of the fast-decaying localized wavelet activation functions in the hidden layer.
- WNNs preserve the universal approximation property, and they are guaranteed to converge with sufficient training.
- WNNs establish an explicit link between the neural network coefficients and the wavelet transform.
- WNNs achieve the same quality of approximation with a network of reduced size.

Designing a WNN requires the researchers to focus particular attention on several areas. First, a suitable learning algorithm is vital in adjusting the weights between the hidden and output layers so that the network does not converge to the undesirable local minima. Second, a proper choice of activation functions in the hidden nodes is crucial as it has been shown that some functions yield significant better result for certain problems [23]. Third, an appropriate initialization of the translation and dilation parameters is essential because this will lead to simpler network architecture and higher accuracy [24].

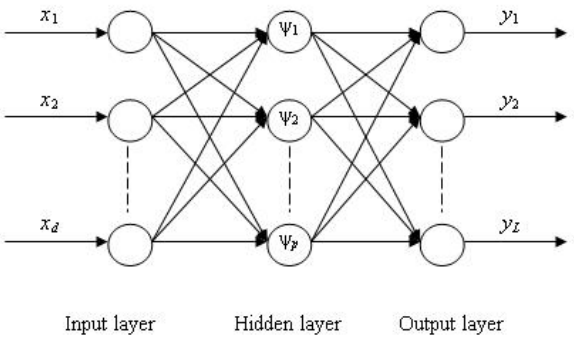


Fig. 3. WNNs with d input nodes, m hidden nodes, and L output nodes.

The selection of the translation vectors for WNNs is of paramount importance. An appropriate initialization of the translation vectors will do a good job of reflecting the essential attributes of the input space, in such a way that the WNNs begin its learning from good starting points and could lead to the optimal solution. Among the notable proposed approaches are the ones given by the pioneers of WNNs themselves, where the translation vectors are chosen from the points located on the interval of the domain of the function [26]. In [10], a dyadic selection scheme realized using the *K*-means clustering algorithm was employed. In [13], the translation vectors were obtained from the new input data. An explicit formula was derived to compute the translation vectors to be used for the proposed composite function WNNs [3]. In [24], an enhanced fuzzy *C*-means clustering algorithm, termed modified point symmetry-distance fuzzy *C*-means (MPSDFCM) algorithm, was proposed to initialize the translation vectors. By incorporating the idea of symmetry similarity measure into the computation, the MPSFCM algorithm was able to find a set of fewer yet effective translation vectors for the WNNs, which eventually led to superb generalization ability in microarray study. In short, the utilization of different novel clustering algorithms in WNNs aim at simpler algorithm complexity and higher classification accuracy from the WNNs.

In this study, the type-2 fuzzy *C*-means (T2FCM) clustering algorithm [16] was proposed to initialize the translation vectors of WNNs. Its clustering effectiveness as well as its robustness to noise has motivated the investigation on the feasibility of T2FCM in selecting the translation vectors of the WNNs. For comparison purposes, the use of *K*-means (KM) and the conventional type-1 fuzzy *C*-means (FCM-1) algorithms in initializing the WNNs translation vectors were also considered.

3.1 Type-2 Fuzzy *C*-Means Clustering Algorithm

Rhee and Hwang [16] proposed an extension to the conventional FCM-1 clustering algorithm by assigning membership grades to type-1 membership values. They pointed out that the conventional FCM-1 clustering may result in undesirable clustering when noise exists in the input data. This is because all the data, including the noise, will be assigned to all the available clusters with a membership value. As such, a triangular membership function is proposed, as shown in the following equation:

$$a_{ij} = u_{ij} - \left(\frac{1 - u_{ij}}{2} \right), \tag{2}$$

where u_{ij} and a_{ij} represent the type-1 and type-2 membership values for input j and cluster center i, respectively. The proposed membership function aims to handle the possible noise that might present in the input data. From Eq. 2, the new membership value, a_{ij}, is defined as the difference between the old membership value, u_{ij} and the area of the membership function, where the length of the base of each of the triangular function is taken as 1 minus the corresponding membership value obtained from FCM-1.

By introducing a second layer of fuzziness, the T2FCM algorithm's concept still conforms to the conventional FCM-1 method in representing the membership values. To illustrate, it can noted from Eq. 2 that a larger value of FCM-1 value (closer to 1) will yield a larger value of T2FCM value as well.

Since the proposed T2FCM algorithm is built upon the conventional FCM-1 algorithm, the formula used to find the cluster centers, c_{ij}, can now be obtained from the following equation that has been modified accordingly, as shown below:

$$c_i = \frac{\sum_{j=1}^{N} a_{ij}^m x_j}{\sum_{j=1}^{N} a_{ij}^m}, \tag{3}$$

where m is the fuzzifier, which is commonly set to a value of 2.

The algorithm for T2FCM is similar to the conventional FCM-1, which aims to minimize the following objective function:

$$J_m(U,V) = \sum_{i=1}^{C}\sum_{j=1}^{N} u_{ij}^m \| x_j - c_i \|^2, \tag{4}$$

but it differs in the extra introduced membership function and also the equation that has been modified to update the cluster centers. In general, the algorithm proceeds as follows:

1. Fix the number of cluster centers, C.
2. Initialize the location of the centers, c_i, $i = 1, 2, \ldots, C$, randomly.
3. Compute the membership values using the following equation:

$$U = [u_{ij}] = \left[\left(\sum_{k=1}^{C} \left(\frac{\| x_j - c_i \|}{\| x_j - c_k \|} \right)^{\frac{2}{m-1}} \right)^{-1} \right]. \tag{5}$$

4. Calculate the new membership value, a_{ij} from the values of u_{ij} using Eq. 2.
5. Update the cluster centers using Eq. 3.
6. Repeat steps 3-5 until the locations of the centers stabilize.

The algorithm for T2FCM is summarized in the flowchart shown in Fig. 4.

3.2 *K*-fold Cross Validation

In statistical analysis, *k*-fold cross validation is used to estimate the generalization performance of classifiers. Excessive training will force the classifiers to memorize the input vectors, while insufficient training will result in poor generalization when a

new input is presented to it. In order to avoid these problems, *k*-fold cross validation is performed.

To implement the *k*-fold cross validation, the samples are first randomly partitioned into *k*>1 distinct groups of equal (or approximately equal) size. The first group of samples is selected as the testing data initially, while the remaining groups serve as training data. A performance metric, for instance, the classification accuracy, is then measured. The process is repeated for *k* times, and thus, the *k*-fold cross validation has the advantage of having each of the sample being used for both training and testing. The average of the performance metric from the *k* iterations is then reported. In this study, *k* is chosen as 10.

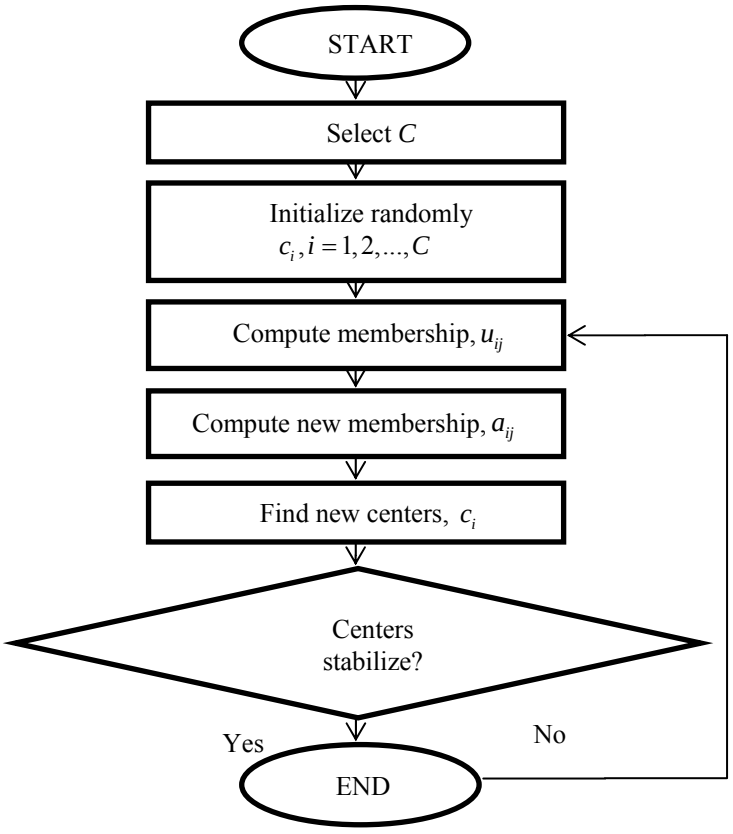


Fig. 4. Algorithm for T2FCM.

4 Results and Discussion

The binary classification task between normal subjects and epileptic patients was realized using the WNNs models. The activation function used in the hidden nodes is the Morlet wavelet function. During the training process, a normal EEG signal was indicated by a single value of 0, while an epileptic EEG signal was labeled with a value of 1. During the testing stage, a threshold value of 0.5 was used, that is, any output from WNNs which is equals to or greater than 0.5 will be reassigned a value of 1; otherwise, it will be reassigned a value of 0. The simulation was carried out using the mathematical software MATLAB® version 7.10 (R2010a). The performance of the proposed WNNs was evaluated using the statistical measures of classification accuracy, sensitivity and specificity. The corresponding classification results between the normal and epileptic EEG signals by using the WNNs-based classifier with different initialization approaches are listed in Table 1.

In terms of the classification accuracy, the translation vectors generated by the conventional KM clustering algorithm gave the poorest result, where an overall accuracy of 94.8% was obtained. The WNNs that used the conventional FCM-1 clustering algorithm reported an overall accuracy of 97.15%. The best performance was obtained by the classifier that employed the T2FCM algorithm, which yielded an overall classification accuracy of 98.87%.

As shown in Table 1, a steady increase in the classification accuracy was noticed when the KM clustering algorithm was substituted with FCM algorithm, and subsequently T2FCM algorithm. FCM outperformed the primitive KM algorithm because the soft clustering employed can assign one particular datum to more than one cluster. On the contrary, KM algorithm, which used hard or crisp clustering, assigns one datum to one center only, and this degrades greatly the classification accuracy. While FCM relies on one fuzzifier, T2FCM adds a second layer of fuzziness by assigning a membership function to the membership value obtained from the type-1 FCM membership values.

In the field of medical diagnosis, the unwanted noise and outliers produced from the signals or images need to be handled carefully, as they will affect and skew the results and analysis obtained afterwards. In this regard, the concept of fuzziness can be incorporated to deal with these uncertainties. Outliers or noise can be handled more efficiently and higher classification accuracy can be obtained via the introduction of the membership function. The noise in the biomedical signals used in this work has thus been handled via two different approaches. The first treatment is in the

Table 1. The performance metrics for the binary classification problem.

Initialization methods	Performance metric		
	Sensitivity	Specificity	Accuracy
KM	85.00	97.30	94.80
FCM	93.82	97.92	97.15
T2FCM	94.96	99.43	98.87

Table 2. Performance comparison of classification accuracy obtained by the proposed WNNs and other approaches reported in the literature

Feature Selection Method	Classifier	Accuracy	References
Time Frequency Analysis	ANNs	97.73	[20]
DWT with KM	MLPs	**99.60**	[15]
DWT	MLPs	97.77	[8]
Approximate Entropy	ANNs	98.27	[8]
This Work		**98.87**	

feature selection stage, where the 10th and 90th percentiles of the absolute values of the wavelet coefficients were used instead of the minima and maxima values. The second way is via the T2FCM clustering algorithm used when initializing the translation parameters for the hidden nodes of WNNs. The clustering achieved by T2FCM proves to result in more desirable locations compared to the conventional KM and FCM-1 methods, as reflected in the higher overall classification accuracy.

Numerous epileptic detection approaches have been implemented in the literature using the same benchmark dataset as in this study. For the sake of performance assessment, comparison of the results with other state-of-the-art methods reported in the literature was included, as presented in Table 2. As depicted in this table, the proposed WNNs with T2FCM initialization approach outperformed the others generally. However, the achieved classification accuracy of 98.87% by the proposed model was inferior to the multilayer perceptrons (MLPs)-based classifier as described in [15], which might be attributed to their feature extraction method. Instead of using basic statistical features, the authors used the KM clustering algorithm to find the similarities among the wavelet coefficient, where the obtained probability distribution from the KM was used as the input of the MLPs-based classifier. A better set of deterministic features might be obtained from this approach, which will be an interesting topic to pursue in future. However, it is pertinent to note that the MLPs-based classifiers are subject to slow learning deficiency and getting trapped in local minima easily.

In order to evaluate the statistical significance of the obtained results, statistical test on the difference of the population mean of the overall classification accuracy was performed using the *t* distribution. The experiment was run 10 times to obtain the values of the summary statistics, namely, the mean and the standard deviation of the samples. The 1% significance level, or α=0.01 was utilized to check whether there is significant difference between the two population means. Two comparisons were done, namely, between KM and T2FCM, and between FCM and T2FCM. The formula for the test statistics is given by:

$$t = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - \left(\mu_1 - \mu_2\right)}{s_{\overline{x}_1 - \overline{x}_2}}, \tag{6}$$

where $\overline{x}_1$ and $\overline{x}_2$ are the sample means; μ_1 and μ_2 are the population means; and $s_{\overline{x}_1 - \overline{x}_2}$ is the estimate of the two standard deviations.

For both cases, the values of the test statistic obtained fall in the rejection region. So the null hypothesis is rejected and it is concluded that there is significant difference between the classification accuracy obtained using the different initialization methods, that is, the performance of T2FCM is superior to those of KM and FCM.

5 Conclusions

In this paper, a novel seizure detection scheme using the improved WNNs with T2FCM initialization approach was proposed. Based on the overall classification accuracy obtained from the real world problem of epileptic seizure detection, it was found that the proposed model outperformed the other conventional clustering algorithms, where an overall accuracy of 98.87%, sensitivity of 94.96% and specificity of 99.43% were achieved. The initialization accomplished via T2FCM has proven that the algorithm can handle the uncertainty and noise in the EEG signals better than the conventional KM and FCM-1 algorithms. This again suggested the prospective implementation of the proposed method in developing a real time automated epileptic diagnostic system with fast and accurate response that could assist the neurologists in their decision making process.

Acknowledgements. The authors gratefully acknowledge the generous financial support provided by Universiti Sains Malaysia under the USM Fellowship Scheme.

References

1. Adeli, H., Zhou, Z., Dadmehr, N.: Analysis of EEG records in an epileptic patient using wavelet transform. J Neurosci Meth 123, 69-87 (2003)
2. Andrzejak, R. G., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C. E.: Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Phys Rev E 64, (2001)
3. Cao, J. W., Lin, Z. P., Huang, G. B.: Composite function wavelet neural networks with extreme learning machine. Neurocomputing 73, 1405-1416 (2010)
4. Ebrahimpour, R., Babakhani, K., Arani, S. A. A. A., Masoudnia, S.: Epileptic Seizure Detection Using a Neural Network Ensemble Method and Wavelet Transform. Neural Netw World 22, 291-310 (2012)
5. Gandhi, T. K., Chakraborty, P., Roy, G. G., Panigrahi, B. K.: Discrete harmony search based expert model for epileptic seizure detection in electroencephalography. Expert Syst Appl 39, 4055-4062 (2012)
6. Ghosh-Dastidar, S., Adeli, H., Dadmehr, N.: Mixed-band wavelet-chaos-neural network methodology for epilepsy and epileptic seizure detection. IEEE transactions on bio-medical engineering 54, 1545-1551 (2007)
7. Ghosh-Dastidar, S., Adeli, H., Dadmehr, N.: Principal component analysis-enhanced cosine radial basis function neural network for robust epilepsy and seizure detection. IEEE transactions on bio-medical engineering 55, 512-518 (2008)

8. Guo, L., Rivero, D., Dorado, J., Rabunal, J. R., Pazos, A.: Automatic epileptic seizure detection in EEGs based on line length feature and artificial neural networks. J Neurosci Meth 191, 101-109 (2010)
9. Guo, L., Rivero, D., Pazos, A.: Epileptic seizure detection using multiwavelet transform based approximate entropy and artificial neural networks. J Neurosci Meth 193, 156-163 (2010)
10. Hwang, K., Mandayam, S., Udpa, S. S., Udpa, L., Lord, W., Atzal, M.: Characterization of gas pipeline inspection signals using wavelet basis function neural networks. NDT and E Int 33, 531-545 (2000)
11. Kandaswamy, A., Kumar, C. S., Ramanathan, R. P., Jayaraman, S., Malmurugan, N.: Neural classification of lung sounds using wavelet coefficients. Comput Biol Med 34, 523-537 (2004)
12. Kumar, S. P., Sriraam, N., Benakop, P. G., Jinaga, B. C.: Entropies based detection of epileptic seizures with artificial neural network classifiers. Expert Syst Appl 37, 3284-3291 (2010)
13. Lin, C.-J.: Nonlinear systems control using self-constructing wavelet networks. Appl Soft Comput 9, 71-79 (2009)
14. Naghsh-Nilchi, A. R., Aghashahi, M.: Epilepsy seizure detection using eigen-system spectral estimation and Multiple Layer Perceptron neural network. Biomed Signal Proces 5, 147-157 (2010)
15. Orhan, U., Hekim, M., Ozer, M.: EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. Expert Syst Appl 38, 13475-13481 (2011)
16. Rhee, F. C. H., Hwang, C. A type-2 fuzzy C-means clustering algorithm. In: Proceedings of the 20th IEEE FUZZ Conference, pp 1926-1929. IEEE Press, New York (2001)
17. Subasi, A.: EEG signal classification using wavelet feature extraction and a mixture of expert model. Expert Syst Appl 32, 1084-1093 (2007)
18. Subasi, A., Gursoy, M. I.: EEG signal classification using PCA, ICA, LDA and support vector machines. Expert Syst Appl 37, 8659-8666 (2010)
19. Tang, Y., Durand, D. M.: A tunable support vector machine assembly classifier for epileptic seizure detection. Expert Syst Appl 39, 3925-3938 (2012)
20. Tzallas, A. T., Tsipouras, M. G., Fotiadis, D. I.: Automatic Seizure Detection Based on Time-Frequency Analysis and Artificial Neural Networks. 2007, (2007)
21. Ubeyli, E. D.: Wavelet/mixture of experts network structure for EEG signals classification. Expert Syst Appl 34, 1954-1962 (2008)
22. Ubeyli, E. D.: Combined neural network model employing wavelet coefficients for EEG signals classification. Digit Signal Process 19, 297-308 (2009)
23. Zainuddin, Z., Ong, P.: Modified wavelet neural network in function approximation and its application in prediction of time-series pollution data. Appl Soft Comput 11, 4866-4874 (2011)
24. Zainuddin, Z., Ong, P.: Reliable multiclass cancer classification of microarray gene expression profiles using an improved wavelet neural network. Expert Syst Appl 38, 13711-13722 (2011)
25. Zandi, A. S., Javidan, M., Dumont, G. A., Tafreshi, R.: Automated Real-Time Epileptic Seizure Detection in Scalp EEG Recordings Using an Algorithm Based on Wavelet Packet Transform. Ieee T Bio-Med Eng 57, 1639-1651 (2010)
26. Zhang, Q. G., Benveniste, A.: Wavelet Networks. Ieee T Neural Networ 3, 889-898 (1992)

Acute Ischemic Stroke Prediction from Physiological Time Series Patterns

Qing Zhang[1,2], Yang Xie[2], Pengjie Ye[1,2], and Chaoyi Pang[2]

[1] Australian e-Health Research Centre/CSIRO ICT Centre
[2] THe University of New South Wales
{qing.zhang,pengjie.ye,chaoyi.pang}@csiro.au
yang.xie@unsw.edu.au

Abstract. Stroke is one of the major diseases that can cause human deaths. However, despite the frequency and importance of stroke, there are only a limited number of evidence-based acute treatment options currently available. Recent clinical research has indicated that early changes in common physiological variables represent a potential therapeutic target, thus the manipulation of these variables may eventually yield an effective way to optimise stroke recovery. Nevertheless the accuracy of prediction methods based on statistical characteristics of certain physiological variables, such as blood pressure, glucose, is still far from satisfactory due to vague understandings of effects and function domain of those physiological determinants. Therefore, developing a relatively accurate prediction method of stroke outcome based on justifiable determinants becomes more and more important to the decision of the medical treatment at the very beginning of the stroke. In this work, we utilize machine learning techniques to find correlations between physiological parameters of stroke patient during 48 hours after stroke, and their stroke outcomes after three months. Our prediction method not only incorporates statistical characteristics of physiological parameters, but also considers physiological time series patterns as key features. Experiment results on real stroke patients' data indicate that our method can greatly improve prediction accuracy to a high precision rate of 94%, as well as a high recall rate of 90%.

Keywords: Stroke, Outcome Prediction, Time Series Data, Machine Learning

1 Introduction

Stroke is a common cause of human death and is a major cause of death after ischemic heart disease [1]. The World Health Organisation (WHO) defines it as "rapidly developing clinical signs of local (or global) disturbance of cerebral function, with symptoms lasting more than 24 hours or leading to death, and with no apparent cause other than of vascular origin" [2]. Recent years research reveals a strong association between physiological homeostasis and outcomes of Acute Ischemic Stroke. Thus understanding determinants of physiological variables, such

as blood pressure, temperature and blood glucose levels, may eventually yield an effective and potentially widely applicable range of therapies for optimising stroke recovery, such as abbreviating the duration of ischaemia, preventing further stroke, or preventing deterioration due to post-stroke complications.

The correlations between blood pressure and stroke outcomes have been widely studied in the literature. It is stated in current guidelines that a significant decrease of BP during the first hours after admission should be avoided, as it correlates with poor outcomes, measured by Canadian Stroke Scale or modified Rankin Score (mRS), at 3 months [10]. Extreme hypertension and hypotension on admission have also been associated with adverse outcome in acute stroke patients [11]. BP values, periodically monitored within the first 72 hours after admission, demonstrate that extreme values still correlate with unfavored outcomes [9]. For example, high baseline of systolic BP is inversely associated with favourable outcome assessed on mRS at 90 days with OR=1.220 and (95% CI: 1.01 to 1.49). Other periodically retrieved statistical properties of BP within 24 hours of ictus, such as maximum, mean, variability etc., have also been investigated. Yong et al. [12] report strong independent association between those properties and the outcome at 30 days after ischemic stroke. For example, variability of systolic BP is inversely associated with favourable outcome with OR=0.57, (95% CI: 0.35 to 0.92).

Research also shows associations between other physiological variables and stroke outcomes. Abnormalities of blood glucose, heart rate variability, ECG and temperature may be predictors of 3-month stroke outcome.

Most of the above analyses are based on periodically recorded physiological parameters, hourly or daily, up to 3 months. Whether continuous data patterns, such as data trends, have a similar predictive role is still uncertain. Although it is clear that the after stroke elevated 24-hours blood pressure levels predict a poor outcome, few studies have investigated the predictive ability of more sophisticate trends, e.g. combined trends of several physiological parameters. Yet this could be an effective way to readily obtain important prognostic information for acute ischemic stroke patients. Dawson et al did pioneering works on associating shorter length (around 10 minutes) beat-to-beat BP with acute ischemic stroke outcomes [8]. They conclude that a poor outcome, assessed by mRS, at 30 days after ischemic stroke is dependent on stroke subtype, beat-to-beat diastolic BP and Mean Arterial Pressure and variability. However in their study, they still use the average values of continuous recordings, instead of time series patterns as predictors. This motivates our research on mining physiological data patterns as effective predictors of acute ischemic stroke outcome.

Obviously mining physiological data patterns can be easily aligned with time series data classification, which is a traditional topic and has attracted intensive studies. Although there exist many sophisticate time series data mining techniques, we find that most of them, if not all, are not applicable to our application scenario, due to the always incomplete, non-isometric physiological data collected from patients. Therefore, in this paper, we incorporate a simple yet powerful time series data pattern analysing method, trend analyses, into

our prediction method. By utilising those trend features, together with values of traditional physiological variables, we design an efficient algorithm that can predict 3-month stroke outcome with high accuracy.

In summary, we list our contributions in this paper:

- We propose using trend patterns of physiological time series data as a new set of stroke outcome prediction features,
- We design a novel prediction algorithm which can accurately predict 3-months stroke outcomes with high precision and recall rate, when tested against a real data set.

The rest of this paper is organised as follows. Section 2 introduces works related to stroke outcome predictions. Section 3 presents our prediction methods. Section 4 reports empirical study results. And section 5 concludes this paper with possible future studies.

2 Related Work

The relationship between beat-to-beat blood pressure (BP) and the early outcome after acute ischemic stroke was firstly described in [8].

A further investigation on BP was done in [6], which investigated detrimental effects of blood pressure reduction in the first 24 hours of acute stroke onset. BP reduction is regarded to have the possibility to worsen an already compromised perfusion in the brain tissue and thus not lowering BP in the early stage after the stroke onset is suggested. However, it lacks further discussion on the relation of higher BP and outcome. Ritter et al. formulated the blood pressure variation by counting threshold violations. Significant difference in the frequency of upper threshold violation occurrences was observed between different time points after stroke [9] . Wong observed some temporal patterns from the changing process of some physiological variables and also attempted to employ such temporal patterns to explain and predict the early outcomes [5]. However, due to the limit of candidate feature set considered in those studies, achieving an accurate prediction is fairly unlikely in those scenarios.

Relationships between other physiological variables and stroke outcome have also been studied in literature. Abnormalities of serum osmolarity, temperature, blood glucose, SPO_2 may be predictors of stroke outcomes. More specifically, heart rate and ECG, can be correlated to stroke outcomes at 3-months:

- Heart Rate Variability: Gujjar et al. reported that heart rate variability is efficient in predicting stroke outcome. Specifically they studied continuous echocardiogram of 25 patients with acute stroke and concluded that the eye-opening score of Glasgow Coma Scale and low-frequency spectral power were factors that were independently predictive of mortality [16].
- ECG: The relationship between ECG abnormalities and stroke outcomes were reported by Christensen et al. They analysed a large cohort of 692 patients and predict that ECG abnormalities are frequent in acute stroke and may conclude 3-month mortality [17].

3 Stroke outcomes prediction

Our prediction method adopts statistical values of physiological parameters and also incorporates the descriptive ability of the physiological patterns as features to predict 3-months stroke outcomes. Particularly, we use the trend pattern of time series data as new add-on features to form an initial feature set. Then we apply the logistic regression method to classify stroke patient outcomes into two groups: good vs. bad. Note that there exist different clinical criteria in defining good/bad outcomes. We will report empirical study results on all criteria in the next section. Cross validation is also adopted to obtain an unbiased assessment of classifier performance, by which the physiological determinants can be accurately identified in the last stage. Finally, we select a subset of features that can most accurately predict 3-months stroke outcomes. Figure 1 presents logic flows of our method. We use Rankin Scale to represent various outcomes at 3 months after stroke (RS3) [18].

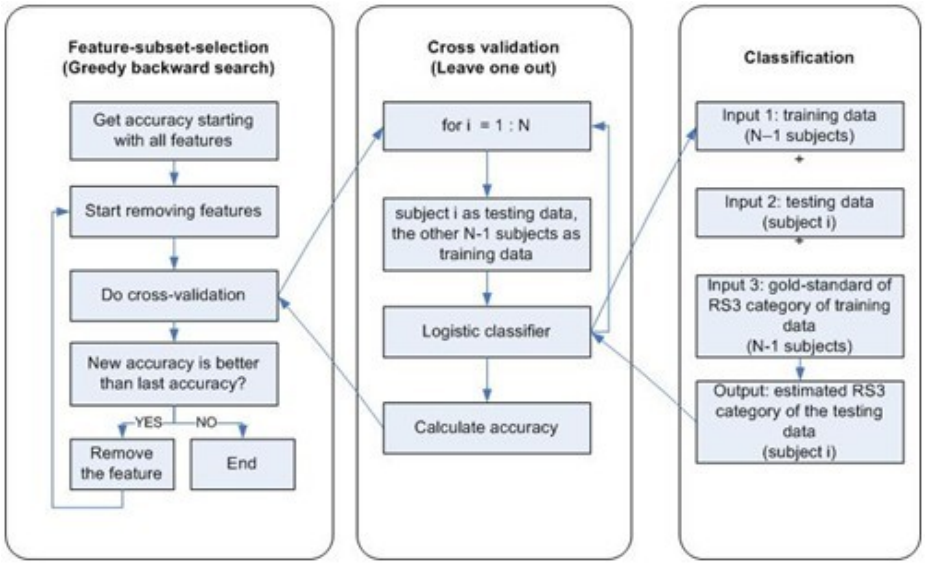


Fig. 1. Stroke outcomes prediction method

3.1 Construct initial feature set

Five physiological parameters are usually considered as influential factors on stroke patient outcomes, namely Blood Sugar Level, Diastolic Blood Pressure, Systolic Blood Pressure, Heart Rate and Body Temperature [6, 16, 17]. Existing stroke outcome predictions always assume a certain parameter as the main

feature in their approaches. However in our approach, we will assume all five parameters in the initial feature set.

Moreover, for each physiological parameter, we compute trends through partitioning the time series data into non-overlapping, continuous blocks. Although there exists many trend and shape detection methods in the literature, such as [3], in our application, we simply consider a bi-partition on the first 48-hours time series data records after stroke. The reasons are:

1. most available physiological data records are only within 48-hours after stroke.
2. clinical observation and our initial experiments both suggest that setting the granularity level at having only two partitions in the 48-hours, well represents the physiological time series pattern changes.

In each partition, accordingly we generate 6 new features, as shown below, to represent the trend pattern:

1. yChange: the difference between the value at the end of a trend and the value at the start of a trend

$$yChange = y(end\_of\_trend) - y(start\_of\_trend)$$

2. absYChange: the absolute value of the yChange
3. slope: the slope of the trend
4. sign: the direction of the trend
5. NumofMeasure: the number of values in a partition
6. FreqofMeasure: the average time interval between measurements, i.e.

$$Freqof Measure = \frac{Trend\_Length}{Numof Measure}$$

The initial feature set comprised physiological values and their trend patterns. We apply the logistical regression method to classify the good/bad stroke outcomes based on this initial feature set.

3.2 Logistic Regression Classifier

In statistics, logistic regression is a type of regression analysis used for predicting the outcome of a binary dependent variable (a variable which can take only two possible outcomes, e.g. “yes” vs. “no” or “success” vs. “failure”) based on one or more predictor variables. Like other forms of regression analysis, logistic regression makes use of one or more predictor variables that may be either continuous or categorical. Unlike ordinary linear regression, however, logistic regression is used for predicting binary outcomes rather than continuous outcomes. Logistic regression adopted here is a type of regression analysis used for predicting the outcome of stroke (“good” vs. “bad”) based on features in our initial feature set.

To obtain an unbiased assessment of classifier performance, the Leave-One-Out Cross validation technique is adopted. Suppose N folds are employed, this

technique withholds a subject from the training set for each run to later test with. Once a record has been withheld for testing, the classifier is trained using the remaining N-1 subjects. The withheld subject is then reintroduced for classification.

3.3 Final feature set selection

We use two greedy search strategies to find the best feature subset that can achieve highest prediction accuracy. Specifically, we use backward search and forward search:

backward search : A greedy backward search is performed to identify a near optimum subset of features. Starting with all features, in sequence, the feature which improves prediction accuracy the most (or decreases it the least) is removed from the current set of features and retained as an intermediate feature subset. This is repeated until all features have been removed. The intermediate feature subset which provides the maximum performance, compared to all other subset evaluated, is selected as the final feature set.

forward Search A sequential forward floating search algorithm is used for feature selection, in an attempt to discover the optimal subset of features from the pool of available candidate features. This strategy begins with a forward-selection process, selecting a single feature from the pool of available features, which improves the prediction accuracy most. After this selection, removal of a feature from the set of selected features is considered. The process of possible feature addition, followed by possible feature removal, is iterated until the selected feature set converges.

4 Empirical Study

In this section, we report experiment results through testing our prediction method on a real data set of stroke patients. Firstly, we introduce the physiological data sets of stroke patients and the good/bad criteria used in our study. Then we report prediction accuracy based on various combination of feature sets. Our study was approved by a ethics committee of the related institution.

4.1 Experimental data sets

A cohort of 157 patients with acute ischaemic stroke were recruited. Patients presenting to the Emergency Department of the Royal Brisbane and Women's Hospital, an Australian tertiary referral teaching hospital, within 48 hours of stroke or existing inpatients with an intercurrent stroke were enrolled prospectively. Important physiological parameters, such as blood pressure, were recorded at least every 4 hours from the time of admission until 48 hours after the stroke.

These values were used as the outcome variable in the analyses. The measurements from patients who died during these first 48 hours were also included in the analyses. Furthermore, some demographic and other stroke-related data were also collected such as the age and gender. The age range of these 157 patients was 16 to 92 years with median age 75 years. The patient distribution based on different values of RS3 is showed in Figure 2.

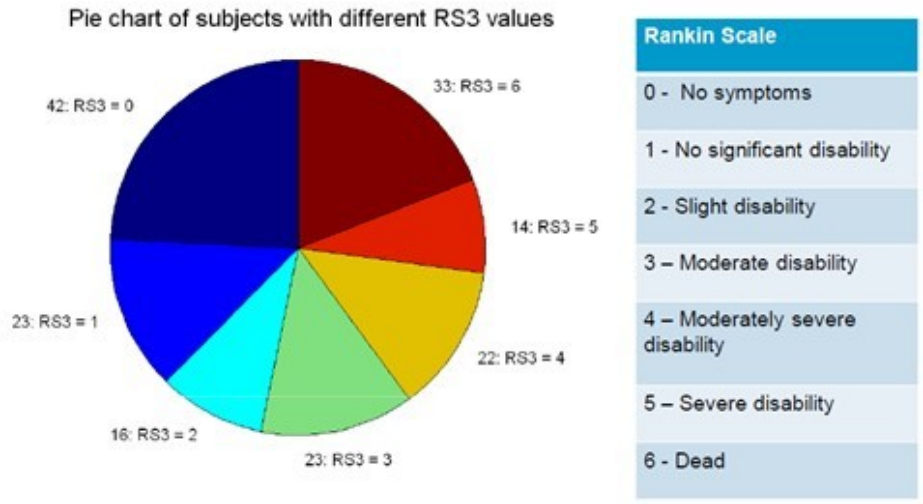


Fig. 2. Patient distributions on values of RS3

4.2 Classification criteria

As shown in Figure 2, RS3 score varies between 0 and 6. Patients with RS3 = 6 means the subject is dead after three months and RS3 = 0 means the subject recovers quite well after three months. Based on RS3 values, patient outcomes can be divided into good/bad groups basing on different grouping criteria. Figure 3 illustrates patient distributions under three type grouping criteria.

4.3 Prediction accuracy comparisons

Applying techniques described in Section 3, we run experiments on various grouping criteria to test our stroke outcome prediction algorithm. We always notice that 'backward search' generates more accurate prediction results, which will thus be used as our default feature set search strategy. Figure 4 shows prediction accuracy comparisons under all three types of grouping criteria. In Figure 5, we also evaluate the efficiency of including trend pattern as prediction

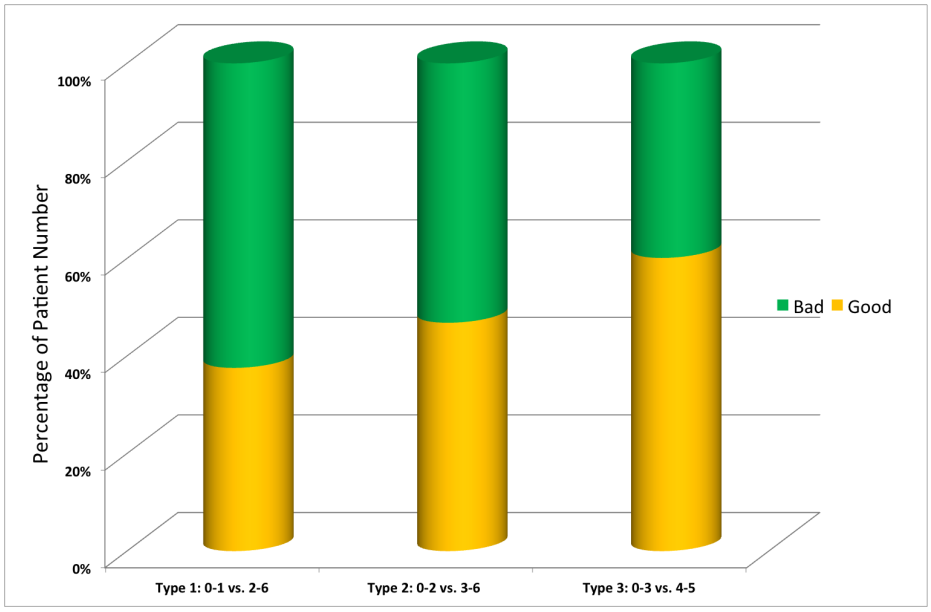


Fig. 3. Good vs Bad outcomes under various criteria

Precision: 86% Recall: 88%		Estimated	
		Good	Bad
True	Good	56	9
	Bad	8	100

Type 1

Precision: 94% Recall: 90%		Estimated	
		Good	Bad
True	Good	76	5
	Bad	9	83

Type 2

Precision: 93% Recall: 88%		Estimated	
		Good	Bad
True	Good	97	7
	Bad	13	56

Type 3

Fig. 4. Prediction Accuracy on different grouping criteria

features. Experiment shows that by adding those simple trend features, the prediction accuracy on all three grouping types is unanimously boosted from 71% to 89∼91%.

5 Conclusion

In this paper, we describe novel algorithms to predict three months stroke outcomes. We have quantified the great improvements brought by including physiological data trend patterns as features of a classifier. We believe that these trends play important roles on three months outcomes of stroke patients. The efficiency and accuracy of our algorithm have also been demonstrated through our experiments.

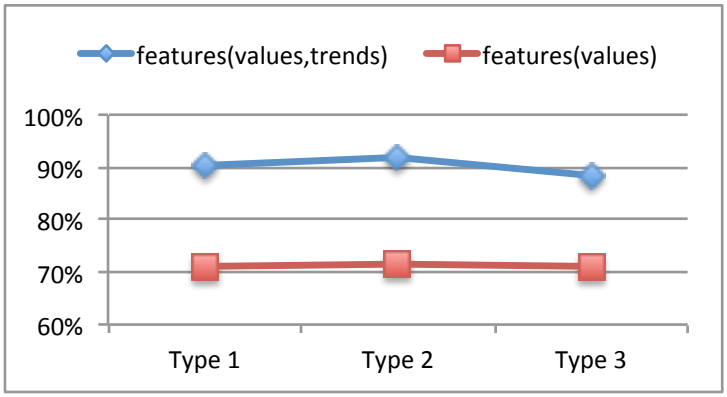


Fig. 5. Prediction accuracy improved by adding trend features

In our future work, we will first try to locate the most important trend pattens for stroke outcome predictions. Then we will work with healthcare professionals to find clinical ground truth beneath those physiological trend patterns of stroke patients. This will greatly benefit clinical treatments of acute ischemic stroke. We also plan to run clinical trials to validate our prediction methods on other real data sets of stoke patients.

References

[1] Australian Institute of Health and Welfare.: Australias health 2006, the tenth biennial health report of the Australian Institute of Health and Welfare. ISBN 1 74024 565 2. 2006

[2] The World Health Organization MONICA Project (monitoring trends and determinants in cardiovascular disease): a major international collaboration. WHO MONICA Project Principal Investigators. Journal of Clinical Epidemiology. 1988;41(2):105-14.

[3] Ye, L., Keogh, E.: Time series shapelets: a new primitive for data mining. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, Vol. 22, ACM, Paris, France, pp. 947–956.

[4] Mueen, A., Keogh, E.,Young, N.: Logical-shapelets: an expressive primitive for time series classification. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, Vol. 22, ACM, San Diego, California, USA, pp. 1154–1162.

[5] Wong, A.: The Natural History and Determinants of Changes in Physiological Variables after Ischaemic Stroke. Ph.D. Thesis, The University of Queensland, St.Lucia.

[6] Oliveira-Filho, J., Silva, S.C.S., Trabuco, C.C., Pedreira, B.B., Sousa, E.U., Bacellar, A.: Detrimental effect of blood pressure reduction in the first 24 hours of acute stroke onset. Neurology. 61(8), 1047–1051.

[7] Marti-Fabregas, J., Belvis, R., Guardia, E., Cocho, D., Munoz, J., Marruecos, L., Marti-Vilalta, J.-L.: Prognostic value of Pulsatility Index in Acute Intracerebral Hemorrhage. Neurology. 61(8), 1051–1056.

[8] Dawson, S.L., Manktelow, B.N., Robinson, T.G., Panerai, R.B., Potter, J.F.: Which Parameters of Beat-to-Beat Blood Pressure and Variability Best Predict Early Outcome After Acute Ischemic Stroke. Stroke. 2000(31), 463–468.
[9] Ritter, M.A., Kimmeyer, P., Heuschmann, P.U., Dziewas, R. Dittrich, R., Nabavi, D.G., Ringelstein, E.B.: Blood Pressure Threshold Violations in the First 24 Hours After Admission for Acute Stroke: Frequency, Timing, Predictors, and Impact on Clinical Outcome. Stroke. 2009(40), 462–468.
[10] Castillo, J., et al., Blood pressure decrease during the acute phase of ischemic stroke is associated with brain injury and poor stroke outcome. Stroke, 2004. 35(2): p.520-6
[11] Ahmed, N., P. Nasman, and N.G. Wahlgren, Effect of intravenous nimodipine on blood pressure and outcome after acute stroke. Stroke, 2000. 31(6): p. 12505.
[12] Yong, M. and M. Kaste, Association of characteristics of blood pressure profiles and stroke outcomes in the ECASSII trial. Stroke, 2008. 39(2): p. 36672
[13] Wong AA, Schluter PJ, Henderson RD, O'Sullivan JD, Read SJ. The natural history of blood glucose within the first 48 hours after ischemic stroke. Neurology 2008;70:103641.
[14] Christensen, H., A. Fogh Christensen, and G. Boysen, Abnormalities on ECG and telemetry predict stroke outcome at 3 months. J Neurol Sci, 2005. 234(12): p. 99103.
[15] Boysen, G. and H. Christensen, Stroke severity determines body temperature in acute stroke. Stroke, 2001. 32 (2): p. 4137.
[16] Gujjar AR, Sathyaprabha TN, Nagaraja D, Thennarasu K and Pradhan N, Heart rate variability and outcome in acute severe stroke: role of power spectral analysis. Neurocrit Care, 2004. 1(3): p. 347-53.
[17] Christensen, H., A. Fogh Christensen, and G. Boysen, Abnormalities on ECG and telemetry predict stroke outcome at 3 months. J Neurol Sci, 2005. 234(1-2): p. 99-103.
[18] Rankin J (May 1957). Cerebral vascular accidents in patients over the age of 60. II. Prognosis. Scott Med J 2 (5): 200-15.

Comparing Data Mining with Ensemble Classification of Breast Cancer Masses in Digital Mammograms

Shima Ghassem Pour[1], Peter Mc Leod[2], Brijesh Verma[2], and Anthony Maeder[1]

[1] School of Computing, Engineering and Mathematics,
University of Western Sydney
Campbelltown, New South Wales, Australia
[2] School of Information and Communication Technology,
Central Queensland University
Rockhampton, Queensland, Australia
{shima.ghassempour,mcleod.ptr}@gamil.com,
b.verma@cqu.edu.au,a.maeder@uws.edu.au

Abstract. Medical diagnosis sometimes involves detecting subtle indications of a disease or condition amongst a background of diverse healthy individuals. The amount of information that is available for discovering such indications for mammography is large and has been growing at an exponential rate, due to population wide screening programmes. In order to analyse this information data mining techniques have been utilised by various researchers. A question that arises is: do flexible data mining techniques have comparable accuracy to dedicated classification techniques for medical diagnostic processes? This research compares a model-based data mining technique with a neural network classification technique and the improvements possible using an ensemble approach. A publicly available breast cancer benchmark database is used to determine the utility of the techniques and compare the accuracies obtained.

Keywords: latent class analysis, digital mammography, breast cancer, clustering, classification, neural network.

1 Introduction

Medical diagnosis is an active area of pattern recognition with different techniques being employed [17, 19, 12]. The expansion of digital information for different cohorts [15] has allowed researchers to examine relationships that were previously not uncovered due to the limited nature of information as well as a lack of techniques being available for the analysis of large data sets. Flexible data mining techniques have the capacity to predict disease and reveal previous unknown trends.

The question that arises is whether the relationships that are revealed by those techniques are as accurate or as comparable as techniques that are specifically developed for other purposes, such as a diagnostic system for a particular

disease or condition. This research aims at contrasting the cluster analysis technique (Latent Class Analysis) of Ghassem Pour, Maeder and Jorm [4] against a baseline neural network classifier, and then considers the effects of applying an ensemble technique to improve the accuracies obtained.

The organisation of this paper is that section two provides a background on the approaches that have been utilised for breast cancer diagnosis, sections three and four detail the proposed techniques for comparison, section five outlines the experimental results obtained and conclusions are presented in section six.

2 Background

Medical diagnosis is a problematic paradigm in that complex relationships can exist in the diagnostic features that are utilised to map to a resultant diagnosis about the disease state. In different cases the state of the disease condition itself can be marked by stages where the diagnostic symptoms or signs can be subtle or different to other stages of the disease. This means that there is often not a clean mapping between the diagnostic features and the diagnosis [13, 14].

Breast cancer screening using mammography provides an exemplar of this situation. Early detection and treatment have been the most effective way of reducing mortality [2] however Christoyianni et al. [1] noted that 10-30% of breast cancers remain undetected while 15-30% of biopsies are cancerous. Taylor and Potts [22] made similar observations in their research. There are many reasons why various cancers can remain undetected. These include the obfuscation of anomalies by surrounding breast tissue, the asymmetry of the breast, prior surgery, natural differences in breast appearance on mammograms, the low contrast nature of the mammogram itself, distortion from the mammographic process and even talc or powder on the outside of the breast making it hard to identify and discriminate anomalies. Even if an anomaly is detected, a high rate of false positives exist [17, 18].

Clustering has provided a widely used mechanism for organising data into similar groupings. The usage of clustering has also been extended to classifiers and detection systems in order to improve detection and provide greater classification accuracy. Kim et al. [9] developed a classifier based on Adaptive Resonance Theory (ART2) where micro-calcifications were grouped into different classes with a three-layered back propagation network performing the classification. The system achieved 90% sensitivity (Az of 0.997) with a low false positive rate of 0.67 per cropped image.

Other researchers such as Mohanty, Senapati and Lenka [16] explored the application of data mining techniques to breast cancer diagnosis. They indicated that data mining medical images would allow for the collection of effective models, rules as well as patterns and reveal abnormalities from large datasets. Their approach was to use a hybrid feature selection technique with a decision tree classifier to classify breast cancer. They utilised 300 images from the MIAS database. They achieved a classification accuracy of 97.7% however their dataset images contained microcalcifications as well as mass anomalies.

3 Latent Class Analysis and Data Mining

Latent Class Analysis (LCA) has been proposed as a mechanism for improved clustering of data over traditional clustering algorithms like k-means [11]. LCA classifies subjects into one of K unobserved classes based on the observed data, where K is a constant and known parameter. These latent or potential classes are then refined based upon their statistical relationships with the observed variables.

LCA is a probabilistic clustering approach: although each object is assumed to belong to one cluster, there is uncertainty about an object's membership of a cluster [11, 10]. This type of approach offers some advantages in dealing with noisy data or data with complex relationships between variables, although as an iterative method there is always some chance that it will be susceptible to noise and in some cases fail to converge.

An advantage of using a statistical model is that the choice of the cluster criterion is less arbitrary. Nevertheless, the log-likelihood functions corresponding to LC cluster models may be similar to the criteria used by certain non-hierarchical cluster techniques [18]. Another advantage of the model-based clustering approach is that no decisions have to be made about the scaling of the observed variables: for instance, when working with normal distributions with unknown variances, the results will be the same irrespective of whether the variables are normalized or not.

Other advantages are that it is relatively easy to deal with variables of mixed measurement levels (different scale types) and that there are more formal criteria to make decisions about the number of clusters and other model features [3]. We have successfully applied LCA for cases in health data mining when the anomalous range of variables results in more clusters than have been expected from a causal or hypothesis based approach [5]. This implies that in some cases LCA may be used to reveal associations between variables that are more subtle and complex.

Unsupervised clustering requires prior specification of the number of clusters K to be constructed, implying that a model for the data is necessary which provides K. The binary nature of the diagnosis problem implies that K=2 should be used in ideal circumstances, but the possibility exists that allowing more clusters would give a better solution (e.g. by allowing several different classes within the positive or negative groups). Consequently a figure of merit is needed to establish that the chosen K value is optimal. In this research the Bayesian Information Criteria (BIC) is determined for the mass dataset in order to gauge the best number of clusters.

Repeated application of the clustering approach can also lead to different solutions due to randomness in starting conditions. In this work we used multiple applications of the clustering calculations to allow improvement in the results, in an ensemble-like approach. Our improvement strategy was based on selection of the most frequent membership of classes per element, over different numbers of clustering repetitions.

4 Neural Network and Ensemble Methods

Neural networks have been advocated for breast cancer detection by many researchers. Various efforts to refine classification performance have been made, using a number of strategies involving some means of choice between alternatives. Ensembles have been proposed as a mechanism for improving the classification accuracy of existing classifiers [6] providing that constituents are diverse.

Zhang et al. [23] partitioned their mass dataset obtained from the DDSM into several subsets based on mass shape and age. Several classifiers were then tested and the best performing classifier on each subset was chosen. They used SVM, k-nearest neighbour and Decision Tree (DT) classifiers in their ensemble and achieved a combined classification accuracy of 72% that was better than any individual classifier.

Surrendiran and Vadivel [21] proposed a technique that could determine what features had the most appropriate correlation on classification accuracy and achieved 87.3% classification accuracy. They achieved this by using ANOVA DA, Principal Component Analysis and Stepwise ANOVA analysis to determine the relationship between input feature and classification accuracy.

Mc Leod and Verma [14] utilised a clustered ensemble technique that relied on the notion that some patterns could be readily identified through clustering (atomic). Other patterns that were not so easily separable (non-atomic) were classified by a neural network. The classification process involved an initial lookup to determine if a pattern was associated with an atomic class however for non-atomic classes a neural network ensemble that had been created through an iterative clustering mechanism (to introduce diversity into the ensemble) was employed. The advantage of this technique is that the ensemble was not adversely affected by outliers (atomic clusters). This technique was applied to the same mass dataset as utilised in this research and achieved a classification accuracy of 91%.

The ensemble utilised in this research was created by fusing together (using the majority vote algorithm) constituent neural networks that were created by varying the number of neurons in the hidden layer to create diverse networks for incorporation into an ensemble classifier.

5 Experimental Results

The experiments were conducted for LCA and neural network techniques and the related ensemble approaches using mass type anomalies from the Digital Database of Screening Mammography (DDSM) [7]. The features used for classification purposes coincided with the Breast Imaging Reporting and Data System (BI-RADS) as this is how radiologists classify breast cancer. The BI-RADS features of density, mass shape, mass margin and abnormality assessment rank are used as they have been proven to provide good classification accuracy [20]. These features are then combined with patient age and a subtlety value [7].

Experiments were performed utilising the clustering technique of Ghassem

Pour, Maeder and Jorm [4] on this dataset. This was achieved using the Latent Gold® software package. The first step was to utilise the analysis feature of LatentGold® to calculate the BIC value and the classification error rate. This information appears in Table 1 below, with Npar designating the resulting parameter value associated with the LCA.

Table 1. LCA Cluster optimisation based on Classification Error.

Clusters	BIC	Npar	Classification Error
2	1238.8	30	0.0303
3	1240.6	38	0.0403
4	1241.8	46	0.0446
5	1254.1	54	0.0470

Minimisation of BIC and the Classification Error determines the best number of clusters for the LCA analysis in terms of classification accuracy and this was found to be 2 clusters. Nevertheless, it might be expected that some further complexity could be identified in higher numbers of clusters, where multiple clusters may exist for either positive or negative classes. The results obtained when cases of more than 2 clusters were merged to form the dominant positive and negative classes, are detailed in Table 2. These results show the instability

Table 2. LCA Classification Technique Accuracy.

Clusters	Accuracy %
2	87.2
3	56.7
4	43.2
5	32.8

of LCA classification for this dataset at higher numbers of clusters, for example the 2-cluster solution gives better accuracy than the 3-cluster solution (merging into 2 clusters) and so forth. From this we conclude that the natural 2-cluster solution is indeed optimal.

In order to provide a comparison, further experiments were performed using a neural network and then applying an ensemble classifier. The neural network and ensemble techniques were implemented in MATLAB® utilising the neural network toolbox. The parameters utilised are detailed in the Table 3 below. Experiments were first performed with a neural network classifier alone, in order to provide a baseline for measuring the classification accuracy on the selected dataset. The results obtained are detailed in Table 4 below. Further experiments were then performed utilising an ensemble technique with a summary of the neural network test results using ten-fold cross validation, as detailed in Table 5 below.

Table 3. Neural network configuration parameters.

Parameter	Value
Hidden Layers	1
Transfer Function	Tansig
Learning Rate	0.05
Momentum	0.7
Maximum Epochs	3000
Root Mean Square Goal	0.001

Table 4. Neural network classification technique accuracy.

Hidden Neurons	Accuracy (%)
13	80
25	80
52	90
111	79

Table 5. NN-ensemble classification technique accuracy.

Networks	Hidden Neurons in Ensemble	Accuracy (%)
6	24,5,15,32,31,43	94
10	24,5,15,32,31,43,50,75,38,59	96.5
13	24,5,15,32,31,43,50,75,38,59,68,79,116	98
15	24,5,15,32,31,43,50,75,38,59,68,79,116,146,14	96

Experiments were also performed for the ensemble-like optimising of results from the LCA technique. It is difficult to match this process directly with the complexity used for the NN-ensemble experiments, so the number of repetitions has been modelled on plausible choice based on dataset size of 100 cases. The results for these experiments are shown in Table 6 below.

Table 6. LCA-ensemble classification technique accuracy.

Repetitions	Accuracy (%)
10	87
20	89
40	91
70	94

6 Discussion and Conclusions

Examination of the results from Tables 1 to 6 demonstrates that the accuracy obtained with the LCA technique is below that of the baseline classification

performed with the neural network. However an ensemble oriented approach enabled improvement of the results from both techniques.

In order to examine the results more closely the sensitivity, specificity and positive predictive value have been calculated for the best performing results for each of the trialled techniques, shown below in Table 7.

Sensitivity is the True Positive diagnosis divided by the True Positive and False Negative components. Sensitivity can be thought of as the probability of detecting cancer when it exists.

Specificity is the True Negative component divided by the True Negative component plus the False Positive component. Specificity can be thought of as the probability of being correctly diagnosed as not having cancer.

Positive Predictive Value (PPV) is the True Positive component divided by the True Positive component plus the False Positive component. PPV is the accuracy of being able to identify malignant abnormalities. The latent class analysis

Table 7. Performance results for the proposed techniques.

Technique	Performance(%)		
	Sensitivity	Specificity	PPV
Latent Class Analysis	80.5	93.9	95.0
LCA-ensemble	82.7	95.2	96.0
Neural Network	91.6	88.4	90.0
NN-ensemble	97.0	97.9	99.0

technique was not as sensitive as the neural network but had better specificity and a higher positive predictive value than the neural network. Both ensemble approaches resulted in substantially better performance, which of course must be traded off against the increased computational cost. The NN-ensemble technique performed the best with good sensitivity, specificity and a high positive predictive value.

The flexibility of clustering techniques such as LCA provides a mechanism for gaining insight from large data repositories. However once patterns in the data become evident it would appear that other less flexible but more specialised techniques could be utilised to obtain analysis at a higher degree of granularity of the data in question.

A summary of the overall performance of the techniques employed in this paper are presented in Figure 1. The optimal LCA-ensemble result, while less than the optimal NN-ensemble result, is obtained with somewhat less processing effort and complexity, and further improvement may be possible.

Future work could look at extending the comparison of LCA with other data mining algorithms to determine their applicability. Breast cancer represents only one problem domain and applying these methods to other datasets would be a logical extension. Our future research will include more experiments with LatentGold® on other breast cancer datasets to determine how different numbers of clusters produce different classification results for a more detailed analysis.

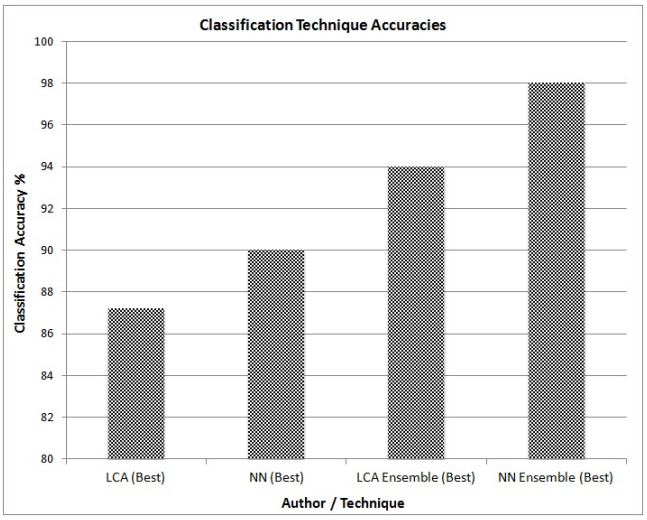


Fig. 1. Comparative Classification Accuracies.

References

1. Christoyianni, I., Koutras, A., Dermatas, E., Kokkinakis, G.: Computer Aided Diagnosis of Breast Cancer in Digitized Mammograms. Computerized Medical Imaging and Graphics 26(5), 309-319 (2002)
2. DeSantis, C., Siegel, R., Bandi, P., Jemal, A.:Breast Cancer Statistics, 2011. CA: A Cancer Journal for Clinicians 61(6), 408-418 (2011)
3. Fraley, C., Raftery, A.: Model-based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association 97(458), 611-631(2002)
4. Ghassem Pour, S., Maeder, A., Jorm, L.: Constructing a Synthetic Longitudinal Health Dataset for Data Mining. DBKDA 2012, The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications.86-90 (2012)
5. Ghassem Pour, S., Maeder, A., Jorm, L.: Validating Synthetic Health Datasets for Longitudinal Clustering. The Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2013) 142, to appear (2013)
6. Gou, S., Yang, H., Jiao, L., Zhuang, X.: Algorithm of Partition Based Network Boosting for Imbalanced Data Classification. The International Joint Conference on Neural Networks (IJCNN).1-6. IEEE (2010)
7. Heath, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P.: The Digital Database for Screening Mammography. Proceedings of the 5th International Workshop on Digital Mammography.212-218 (2000)
8. Hofvind, S., Ponti, A., Patnick, J., Ascunce, N., Njor, S., Broeders, M., Giordano, L., Frigerio, A., Tornberg, S.: False-positive Results in Mammographic Screening for Breast Cancer in Europe: a literature review and survey of service screening programmes. Journal of Medical Screening 19(1), 57-66 (2012)
9. Kim, J., Park, J., Song, K., Park, H.: Detection of Clustered Microcalssifications on Mammograms Using Surrounding Region Dependence Method and Artificial Neural Network. The Journal of VLSI Signal Processing 18(3),251-262 (1998)

10. Lanza, S., Flaherty, B., Collins, L.: Latent Class and Latent Transition Analysis. Handbook of Psychology. 663-685 (2003)
11. Magidson, J., Vermunt, J.: Latent Class Models for Clustering: A Comparison with k-means. Canadian Journal of Marketing Research 20(1), 36-43 (2002)
12. Malich, A., Schmidt, S., Fischer, D., Facius, M., Kaiser, W.: The Performance of Computer-aided Detection when Analyzing Prior Mammograms of Newly Detected Breast Cancers with Special Focus on the Time Interval from Initial Imaging to Detection. European Journal of Radiology 69(3),574-578 (2009)
13. Mannila, H.: Data mining: Machine learning, Statistics, and Databases. Proceedings of Eighth International Conference on Scientific and Statistical Database Systems.2-9 IEEE (1996)
14. McLeod, P., Verma, B.: Clustered Ensemble Neural Network for Breast Mass Classiffcation in Digital Mammography. In: The International Joint Conference on Neural Networks (IJCNN). 1266-1271 (2012)
15. Mealing, N., Banks, E., Jorm, L., Steel, D., Clements, M., Rogers, K.: Investigation of Relative Risk Estimates from Studies of the Same Population with Contrasting Response rates and Designs. BMC Medical Research Methodology 10(1), 10-26 (2010)
16. Mohanty, A., Senapati, M., Lenka, S.: A Novel Image Mining Technique for Classification of Mammograms Using Hybrid Feature Selection. Neural Computing & Applications. 1-11 (2012)
17. Nishikawa, R., Kallergi, M., Orton, C., et al.: Computer-aided Detection, in its present form, is not an Effective aid for Screening Mammography. Medical Physics 33(4), 811-814 (2006)
18. Nylund, K., Asparouhov, T., Muthen, B.: Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. Structural Equation Modeling 14(4), 535-569 (2007)
19. Oh, S., Lee, M., Zhang, B.: Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics 8(2), 316-325 (2011)
20. Sampat, M., Bovik, A., Markey, M.: Classification of Mammographic lesions into BIRADS Shape Categories Using the Beamlet Transform. In: Proceedings of SPIE, Medical Imaging: Image Processing. 16-25. SPIE(2005)
21. Surrendiran, B., Vadivel, A.: Feature Selection Using Stepwise ANOVA, Discriminant Analysis for Mammogram Mass Classification. International Journal of Recent Trends in Engineering and Technology 3, 55-57 (2010)
22. Taylor, P., Potts, H.: Computer Aids and Human Second Reading as Interventions in Screening Mammography: two systematic reviews to compare effects on cancer detection and recall rate. European Journal of Cancer 44(6), 798-807 (2008)
23. Zhang, Y., Tomuro, N., Furst, J., Raicu, D.: Building an Ensemble System for Diagnosing Masses in Mammograms. International Journal of Computer Assisted Radiology and Surgery 7(2), 323-329 (2012)

Automatic Classification of Cancer Notifiable Death Certificates

Luke Butt[1], Guido Zuccon[1], Anthony Nguyen[1],
Anton Bergheim[2], Narelle Grayson[2]

[1] The Australian e-Health Research Centre, Brisbane, Queensland, Australia;
[2] Cancer Institute NSW, Alexandria, New South Wales, Australia.
{luke.butt, guido.zuccon, anthony.nguyen}@csiro.au
{anton.bergheim, narelle.grayson}@cancerinstitute.org.au

Abstract. The timely notification of cancer cases is crucial for cancer monitoring and prevention. However, the abstraction and classification of cancer from the free-text of pathology reports and other relevant documents, such as death certificates, are complex and time-consuming activities. In this paper we investigate approaches for the automatic detection of cases where the cause of death is a notifiable cancer from free-text death certificates supplied to Cancer Registries. A number of machine learning classifiers were investigated. A large set of features were also extracted using natural language techniques and the Medtex toolkit; features include stemmed words, bi-grams, and concepts from the SNOMED CT medical terminology. The investigated approaches were found to be very effective in identifying death certificates where the cause of death was a notifiable cancer. Best performance was achieved by a Support Vector Machine (SVM) classifier with an overall F-measure of 0.9647 when evaluated on a set of 5,000 free-text death certificates. This classifier considers as features stemmed token bigrams and information from SNOMED CT concepts filtered by morphological abnormalities and disorders. However, our analysis shows that it is the selection of features that most influences the performance of the classifiers rather than the type of classifier or the feature weighting schema. Specifically, we found that stemmed token bigrams with or without SNOMED CT concepts are the most effective feature. In addition, the combination of token bigrams and SNOMED CT information was found to yield the best overall performance.

Keywords: death certificates, Cancer Registry, cancer monitoring and reporting, machine learning, natural language processing, SNOMED CT

1 Introduction

Cancer notification and reporting is an important and fundamental process for providing an accurate picture of the impact of cancer, the nature and extent of cancer, and to direct research efforts for the cure of cancer. Cancer Registries collect and interpret data from a large number of sources, helping to improve cancer

prevention and control, as well as treatments and survival rates for patients with cancer.

The manual coding of documents, such as pathology reports and death certificates, with respect to notifiable cancers and corresponding synoptic factors (such as primary site, morphology, etc.) is a laborious and time consuming process. Cancer Registries strive to provide timely and accurate information on cancer incidence and mortality in the community. They receive large quantities of data from a range of sources, including hospitals, pathology laboratories and Registries of Births, Deaths and Marriages (which issues releases of death certificates). It is estimated that incident cases within Cancer Registries that have death certificate only notifications amount to about 1-5% of the total cases; delays in the processing of this data may cause underestimation of the incidence of cancer. Computational methods for the automatic abstraction of relevant information have the possibility to enhance a Cancer Registry's workflow, providing time and costs savings as well as timely cancer incidence information and mortality information. This automatic process is however challenging, both for the complex nature of the language used in the reports, and for the high level of recall and accuracy required.

Previous works have attempted to provide automatic cancer coding from free-text pathology reports collected by Cancer Registries. For example, Nguyen et al. [1] used natural language processing techniques and a rule-based system to automatically extract relevant synoptic factors from electronic pathology reports. Similarly, Zuccon et al. [2] showed how these techniques could cope with character recognition errors generated by scanning free-text pathology reports stored in paper form. Machine learning approaches have also been considered; for instance, D'Avolio et al. [3] have tested approaches based on supervised machine learning (Conditional Random fields and Maximum Entropy) and have shown its effectiveness for the classification of pathology reports that were consistent with cancer in the domains of colorectal, prostate, and lung cancer.

Cancer Registries have access to a number of data sources beyond pathology reports. One such data source is death certificates. Death certificates are a rich source of data that can support cancer surveillance, monitoring and reporting. These certificates contain free-text sections that report the cause of the death of an individual. An example of the free-text content of a death certificate where the cause of death is a notifiable cancer is given in Figure 1, while Figure 2 is an example of a non-notifiable death certificate.

Limited works have focused on computational methods for automatically classifing death certificates with respect to the cause of death. The SuperMICAR system and its related tools[1] provide a semi-automatic coding of the cause of death in death certificates. The system identifies keywords and expressions from the free-text documents that indicate possible causes of death; this is done through the use of a standard set of expressions encoded in a predefined vocabulary. Extracted free-text expressions are then converted to one or more

[1] Consult `http://www.cdc.gov/nchs/nvss/mmds/super_micar.htm` (last visited 19th November 2012) for further details.

```
(I)A) MAXILLARY TUMOR, 2 YEARS B) PULMONARY OEDEMA, 1 WEEK
(II) CEREBROVASCULAR ACCIDENT/DYSPLASIA, 20 YEARS ASTHMA
```

Fig. 1. A de-identified death certificate where the cause of death is a notifiable cancer.

```
I(A) CEREBROVASCULAR ACCIDENT 48 HOURS (B) CEREBRAL ARTERIOSCLEROSIS YEARS
(C) HYPERTENSION YEARS II CHRONIC ALCOHOLISM YEARS
```

Fig. 2. A de-identified death certificate where the cause of death is not a notifiable cancer.

ICD-10 codes which are then aggregated into a single ICD-10 underlying cause of death through the use of a rule-base. While doctor reported death certificates can be fed directly into the system, Coroner reported ones require additional pre-processing. A consistent number (between 15 and 20 percent according to a US study [4]) of death certificates cannot be coded through SuperMICAR and related tools, and thus require manual coding. A recent work has successfully classified death certificates related to pneumonia and influenza using a natural language processing pipeline and rule-based system [5]. However, to the best of our knowledge, no previous research has been conducted to investigate fully automatic methods that go beyond keyword spotting of standard cause of death expressions to classifying death certificates, in particular focusing on certificates where the main cause of death is cancer. Furthermore, while Australian Cancer Registries can acquire free-text death certificates on a fortnightly basis from the Registry of Births Deaths and Marriages, coded causes of death produced by SuperMICAR (and related products) are released by the Australian Bureau of Statistics on a yearly basis. Computational methods able to tackle the fast identification of death certificates where the cause of death is a notifiable cancer would enhance the cancer reporting and monitoring capabilities of Cancer Registries.

In this paper, we focus on the problem of automatically identifying death certificates where the main cause of death is cancer. This problem is cast into a binary classification problem, i.e. death certificates are classified as containing a death cause related to cancer or vice versa as not containing a death cause related to cancer. Several machine learning classifiers were investigated for this task. These include support vector machine, Naive Bayes, decision trees, and boosting algorithms. A state-of-the-art information extraction tool (Medtex [6]) is used to create different set of features that are used to train the classifiers; different feature weighting schemas were also considered. Features include stemmed tokens, n-grams, as well as SNOMED CT concept ids and tokens from fully specified names of SNOMED CT concepts, among others. SNOMED CT is a medical terminology which formally describes in detail the coverage and knowledge of topics and terminology used in the medical domain.

Our approaches are tested on 5,000 de-identified death certificates acquired from an Australian Cancer Registry, using 10-fold cross validation for allowing robust training and testing. Our experimental results demonstrate that the

choice of classifier and weighting schema, although being important, is not critical for achieving high classification effectiveness. Instead, the choice of features used to represent content of death certificates is the determining factor for high classification effectiveness. Specifically, stemmed token bigrams are found to be the single most important features among those extracted. Furthermore, we found that SNOMED CT features provide consistent increments in classification effectiveness if used along with stemmed token bigrams; although not providing a large increment, the combined use of stemmed token bigrams and SNOMED CT morphology provide the best classification effectiveness in our experiments.

Next, we detail the approaches adopted in this paper. Then, in Section 3 we outline our empirical evaluation methodology; classification results obtained by the investigated approaches are reported in Section 4. An analysis of the results is developed in Section 4.1. The paper concludes in Section 5 summarising our main contribution and directions for future work.

2 Approaches for Automatic Classification of Death Certificates

In this paper we investigate supervised machine learning approaches for the detection of death certificates where the cause of death is a notifiable cancer. These approaches are characterised by three main variables: (1) the features extracted from the documents (Section 2.1), (2) the weighting schemas applied to the features to represent documents (Section 2.2), and (3) the specific binary classifier used to individuate certificates where the cause of death is a notifiable cancer (Section 2.3).

2.1 Automatic Feature Extraction

Machine learning algorithms require data to be represented by features, such as the words that occur in a text document. We used the information extraction capabilities of the Medtex system[2] for obtaining a set of meaningful features from the free-text of the death certificates.

The feature sets investigated in this paper are:

stem: a token stem, i.e. the stemmed version of a word contained in a certificates

stemBigram: the bi-gram formed by two token stems, i.e. a pair of adjacent stemmed words as found in a certificates

concept: SNOMED CT concepts as found in the free-text of the certificates using the Medtex system

conceptFull: the tokens of the fully specified name of the extracted SNOMED CT concepts

[2] Medtex comprises both information extraction capabilities (extracting both low level information such as word tokens and stems, punctuation, etc., and higher level semantic information such as UMLS and SNOMED CT concepts [1]) and classification capabilities integrated via its rule-based engine.

concFullMorph: the tokens of the fully specified name of extracted SNOMED CT concepts that are morphologic abnormalities or disorders

concBigram: the bigram formed by two adjacent SNOMED CT concept ids

concFullBigram: the bigram formed by two adjacent tokens in the fully specified name of concepts extracted from SNOMED CT

While features like stem and stemBigram are commonly used for classifying free-text documents, features based on SNOMED CT concepts and its properties such as tokens from the fully specified name have not been exploited by previous works that attempted to classify free-text death certificates. SNOMED CT provides a standard clinical terminology used to map various descriptions of a clinical concept to a single standard clinical concept. In this work, the SNOMED CT ontology was used as an underlying mechanism to classify free-text using semantically matching SNOMED CT concepts.

In addition, we also considered pair-wise combinations of features that provided promising results on preliminary experiments. In this paper we shall report the results obtained by all features used singularly, and of the combinations concept + stem, concept + stemBigram, concFullMorph + stemBigram, and concBigram + stemBigram, which has shown promise in preliminary investigations.

Next, we consider the example death certificates given in Figure 1 and Figure 2 to describe how a feature set is constructed. To build the feature representations, we examine each death certificate and for each occurring instance of a feature in the certificate we assign a value of 1, while the absence of a feature is marked by a zero entry value. Note that these values are subsequently modified according to the feature weighting functions, as we shall describe in Section 2.2. After all certificates have been processed in this manner, we add a final feature cancerNotifiable, whose value is obtained from ground truth judgements supplied with the data. Table 1 shows an extract of the feature data constructed for the two example death certificates. The task of the machine learning classifiers is to predict the value of the cancerNotifiable feature, given the learning data supplied.

	Features																			
	stem						stemBigram					concept			conceptFull				...	
Document	ACCID	ALCOHOL	...	TUMOR	WEEK	YEAR	ACCID_DYSPLASIA	ACCID_48	...	20_YEAR	YEAR_ASTHMA	126550004	...	230690007	Neoplasm of maxilla	...	Cerebrovascular accident	Cerebral arteriosclerosis	...	cancerNotifiable
Figure 1	1	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	1	0	...	1
Figure 2	1	1	1	0	0	1	0	1	1	0	0	0	1	1	0	1	1	1	...	0

Table 1. Feature data built from two example death certificates.

Note that no further processing is applied to the text, for example, for removing punctuation, identifying section or list labels, or for removing or correcting typographical errors present in the free-text. While adequate text pre-processing may enhance the quality of the text itself and thus of the extracted features, we left this for future work and instead we focused on investigating weighting schemas for the selected features and binary classifiers.

2.2 Feature Weighting

A number of weighting schemes for capturing the local importance of a feature in a report were tested.

Binary coefficients were used to encode the presence or absence of a feature. We refer to this schema as binary.

The weighting schema composed by the feature frequency $f(\mathcal{F})$ of feature $\mathcal{F}$ was used to capture the number of times a specific feature appeared within a document. We shall refer to this weighting schema as frequency.

Variations of the frequency weighting schema were also experimented with. In this weighting schema, features frequencies were directly translated into weights, i.e. weights are linearly derived from frequencies. Variations consider non-linear functions of the frequency of a feature.

A first variation was to scale the appearance of feature $\mathcal{F}$ in a free-text death certificate by the function $1 + \log(f(\mathcal{F}))$ if $f(\mathcal{F}) \geq 1$, and 0 if the feature was absent. This function would capture the fact that little importance is given to subsequent appearances of a feature $\mathcal{F}$ in a document: the logarithm of a number greater than one plateaus rapidly. In the following, we shall refer to this weighting schema as LogF, i.e. logarithm of the frequency.

A second variation was to assign increasing weights to features that appear with high frequencies within the death certificate. To this aim, the appearance of feature $\mathcal{F}$ was weighted according to the function $e^{f(\mathcal{F})}$, while a zero value was assigned to absent features. It is suggested that, given the short length of the considered death certificates, the unexpected multiple occurrence of a feature would provide strong evidence that that feature is important for the document. Using the exponential function to weight occurrences of a feature would assign dominating scores to features that occur frequently in a document. We shall refer to this weighting function as expF.

Note that only local weighting functions were used to assign scores to features,that is, weights were computed only by taking into account the frequencies of appearance of a feature in a text, thus ignoring the distribution of that feature on a global level, i.e. across the dataset. The incorporation of global occurrence statistics within the weighting schemas is left to future work.

2.3 Automatic Classification Methodology

A number of common classifiers were evaluated. These comprised statistical models (Naive Bayes), support vector machines (SPegasos), decision trees (C4.5), and

boosting algorithms (AdaBoost). We considered the implementations of these algorithms provided in the Weka toolkit [7].

The multinomial Naive Bayes classifier determines the class of a death certificate according to the features that occur in the text and their weights. The SPegasos classifier uses a stochastic gradient descent algorithm and a hinge loss function to produce the separation hyperplane used by the linear support vector machine. In the C4.5 classifier, information gain is used for choosing at each level of the decision tree the most effective feature able to split the data into the two binary classes considered here (i.e. death certificates related to cancers and those not related to cancer). Adaboost minimises of a convex loss function built from the prediction of a base weak classifier. A simple binary decision tree classifier that constructs one-level trees was used as base classifier for Adaboost.

Parameters of all classifiers were set to the default values described in Witten et al. [7].

3 Experimental Methodology

3.1 Data

A set of 5,000 free-text death certificates was acquired from Cancer Institute NSW, the institutional entity responsible for maintaining the Central Cancer Registry in New South Wales. Ethics approval was granted by the NSW Population & Health Services Research Ethics Committee for this study including to use the de-identified data. The free-text documents were short in length, containing on average 13.08 words; the (unstemmed) vocabulary contained 3,751 unique words (including section headings and labels).

Cause of death classifications based on ICD-10 codes accompanied the reports. This coding set was acquired from the Australian Bureau of Statistics, who releases coded data yearly. ICD-10 codings were used to determine the class each death certificates belonged to. A list of ICD-10 codes that are cancer notifiable was provided by Cancer Institute NSW.

The 5,000 death certificates were extracted from Cancer Institute NSW archives so that documents were uniformly split across the two classes, i.e. 2,500 certificates were coded with ICD-10 codes that are for notifiable cancers according to the business rules of Cancer Institute NSW, while the remaining 2,500 were not cancer notifiable. The causes of death of the 2,500 death certificates for notifiable cancers span a total of 367 unique ICD-10 codes.

3.2 Evaluation

A 10-fold cross validation methodology was used to train and test the classification algorithms. In this methodology, the dataset was randomly divided into 10 stratified[3] folds of equal dimensions. A model for each classifier was then learnt

[3] Folds were automatically stratified with respect to the two target classes, not the ICD-10 codes.

on nine of these folds, leaving one fold out for testing. The process was repeated by selecting a new fold for testing, while a new model was learnt from the remaining folds. Classification effectiveness was then averaged across the folds left out for testing in each iteration.

F-Measure (F-m) was used as primary metric to evaluate the efficacy of the implemented classifiers; accuracy, recall (sensitivity, Rec) and precision (positive predictive value, Prec) were also recorded, along with the number of true positive (TP), false positve (FP), true negative (TN), and false negative (FN) classifications.

4 Results and Discussion

The combination of 10 features, 4 weighting schemas, and 4 classifiers requires the evaluation of a total of 160 classifier settings (referred to as runs in the following) on the dataset consisting of 5,000 death certificates. While we evaluated all combinations of features, weighting schema and classifiers, given the large number of combinations, it is not feasible to report the individual results for each of the runs. Thus, we report only the settings of the 40 most effective runs in terms on F-measure, our primary evaluation metric (Table 2), with the F-measure of each classifier over all experimented settings graphically shown in Figure 3. Later in the paper we shall consider a summary evaluation of the variability of results provided by features, weighting schemas, and classifiers. This analysis will comprise of the results from all runs.

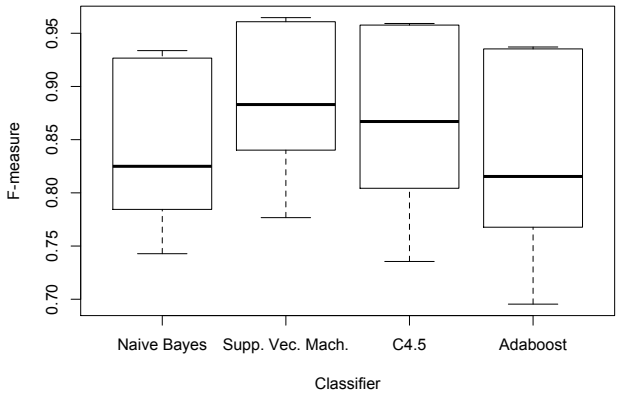


Fig. 3. Boxplot summarising the F-measure performance of the investigated classifiers over all considered settings.

The results reported in Table 2 suggest that the tested approaches are highly effective in discriminating between those death certificates that contain a cancer notifiable cause of death and those that do not.

Classifier	Feature	Weight	Prec	Rec	F-m	TP	FN	FP	TN
SPegasos	concFullMorph + stemBigram	frequency	.9794	.9504	**.9647**	2376	124	50	2450
SPegasos	concFullMorph + stemBigram	logF	.9786	.9500	.9641	2375	125	52	2448
SPegasos	concept + stemBigram	logF	.9770	**.9508**	.9637	**2377**	**123**	56	2444
SPegasos	concFullMorph + stemBigram	binary	.9770	.9504	.9635	2376	124	56	2444
SPegasos	concept + stemBigram	binary	.9766	.9504	.9633	2376	124	57	2443
SPegasos	concept + stemBigram	frequency	.9766	.9504	.9633	2376	124	57	2443
SPegasos	stemBigram	binary	.9761	.9488	.9623	2372	128	58	2442
SPegasos	concFullMorph + stemBigram	expF	.9773	.9476	.9622	2369	131	55	2445
SPegasos	stemBigram	logF	.9753	.9476	.9612	2369	131	60	2440
SPegasos	stemBigram	expF	.9785	.9444	.9611	2361	139	52	2448
SPegasos	stemBigram	frequency	.9764	.9452	.9606	2363	137	57	2443
SPegasos	concept + stemBigram	expF	.9741	.9460	.9598	2365	135	63	2437
C4.5	concept + stemBigram	logF	.9800	.9392	.9592	2348	152	48	2452
C4.5	concept + stemBigram	expF	.9800	.9392	.9592	2348	152	48	2452
C4.5	concept + stemBigram	frequency	.9800	.9392	.9592	2348	152	48	2452
C4.5	concept + stemBigram	binary	.9799	.9384	.9587	2346	154	48	2452
C4.5	concFullMorph + stemBigram	logF	.9856	.9324	.9583	2331	169	34	2466
C4.5	concFullMorph + stemBigram	expF	.9856	.9324	.9583	2331	169	34	2466
C4.5	concFullMorph + stemBigram	frequency	.9856	.9324	.9583	2331	169	34	2466
C4.5	stemBigram	logF	.9848	.9320	.9577	2330	170	36	2464
C4.5	stemBigram	expF	.9848	.9320	.9577	2330	170	36	2464
C4.5	stemBigram	frequency	.9848	.9320	.9577	2330	170	36	2464
C4.5	concFullMorph + stemBigram	binary	.9848	.9320	.9577	2330	170	36	2464
C4.5	stemBigram	binary	.9848	.9308	.9570	2327	173	36	2464
AdaBoost	concept + stemBigram	binary	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concept + stemBigram	logF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concept + stemBigram	expF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concept + stemBigram	frequency	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	binary	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	logF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	expF	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	concFullMorph + stemBigram	frequency	**1**	.8816	.9371	2204	296	**0**	**2500**
AdaBoost	stemBigram	binary	**1**	.8784	.9353	2196	304	**0**	**2500**
AdaBoost	stemBigram	logF	**1**	.8784	.9353	2196	304	**0**	**2500**
AdaBoost	stemBigram	expF	**1**	.8784	.9353	2196	304	**0**	**2500**
AdaBoost	stemBigram	frequency	**1**	.8784	.9353	2196	304	**0**	**2500**
SPegasos	stem	logF	.9588	.9120	.9348	2280	220	98	2402
SPegasos	stem	frequency	.9611	.9096	.9346	2274	226	92	2408
Naive Bayes	stemBigram	binary	.9658	.9036	.9337	2259	241	80	2420
Naive Bayes	concept + stemBigram	binary	.9606	.9076	.9334	2269	231	93	2407

Table 2. Top 40 results with respect to decrease F-measure (F-m).

Overall, the best classifier is the support vector machine implementation provided by SPegasos when used on concFullMorph + stemBigram features, i.e. the fully specified names of concepts associated to morphological abnormalities and disorders as encoded in SNOMED CT, weighted using raw frequencies. SPegasos is found to be very effective also when other combinations of weighting schemas

and features are considered. In addition, this support vector machine classifier shows the smallest variance across all considered settings (Figure 3).

Among the best performing classifiers, AdaBoost used in conjunction with stemmed bigrams features achieved perfect precision (Prec= 1), at the expense of recall. Although these results are remarkable, high precision may be considered less important than high recall in such task. In fact, in a Cancer Registry setting, it is preferable to have high recall and be considering death certificate that are incorrectly reported as containing cancer notifiable cause of death, than to have missed cancer cases. This becomes particularly important if the missed cancer cases refer to rare cancers. AdaBoost also exhibits the highest variance across experiment settings among the considered classifiers (see Figure 3).

4.1 The Impact of Classifiers, Weighting Schemas, and Features

To better understand the role of specific features, weighting schema, and classifiers on the effectiveness of the tested approaches, an analysis of the empirical results where each of the three key characteristics were treated as the controlled variable is performed.

We start by examining the impact of each classification model on the overall effectiveness of the approaches. Table 3 reports maximum (*Max(F-m)*), minimum (*Min(F-m)*), difference (Δ), and variance of F-measure over all runs of each classifier model. SPegasos is found to be the classifier achieving the highest maximum and minimum F-measure values, thus extending the observations made on this classifier when examining the results of Table 2. Instead, while the Naive Bayes classifier was not found to be amongst the most effective classification models in our experiments, its robustness is second only to that of SPegasos, with performance ranges between 0.9337 and 0.7428 in F-Measure. While models such as C4.5 and Adaboost achieve higher values of F-measure than Naive Bayes, their minimum performances are lower than that recorded for Naive Bayes.

Classifier	Max(F-m)	Min(F-m)	Δ	Variance
SPegasos	**0.9647**	**0.7767**	**0.1880**	$\mathbf{5.10 \cdot 10^{-3}}$
Naive Bayes	0.9337	0.7428	0.1909	$\mathbf{5.10 \cdot 10^{-3}}$
C4.5	0.9592	0.7355	0.2237	$7.35 \cdot 10^{-3}$
AdaBoostM1	0.9371	0.6954	0.2417	$7.88 \cdot 10^{-3}$

Table 3. Classification effectiveness across the four classifiers ordered by increasing max-min F-measure range (Δ).

We continue by analysing the influence of weighting schemas on the classification results of the approaches investigated in this work. Simple raw frequency weighting, i.e. frequency, is found to be the most effective weighting schema. However, no weighting schema appears to be significantly better than another: while

Weight	Max(F-m)	Min(F-m)	Δ	Variance
binary	0.9635	0.6954	0.2681	$6.81 \cdot 10^{-3}$
frequency	**0.9647**	0.6954	0.2693	$6.74 \cdot 10^{-3}$
logF	0.9641	0.6954	0.2687	$6.80 \cdot 10^{-3}$
expF	0.9622	0.6954	**0.2668**	$\mathbf{6.53 \cdot 10^{-3}}$

Table 4. Classification effectiveness across the four weighting schema ordered by increasing max-min F-measure range (Δ).

frequency achieves the best performance with a F-measure of 0.9647, the highest F-measure of the worst performing schema is 0.9622 (expF), just 0.003% lower than frequency. Furthermore, all weighting schema exhibit the same effectiveness when considering the worst performing settings. Thus the range of performance differences and their variance do not significantly differ across weighting schema. This may be due to the fact that death certificates are in general short documents, where features occur uniformly.

Feature	Max(F-m)	Min(F-m)	Δ	Variance
stemBigram	0.9623	**0.9275**	**0.0348**	$\mathbf{2.02 \cdot 10^{-4}}$
concept + bigramStem	0.9637	0.9267	0.0370	$2.16 \cdot 10^{-4}$
concFullMorph + stemBigram	**0.9647**	0.9255	0.0392	$2.33 \cdot 10^{-4}$
concBigram + stemBigram	0.8443	0.7677	0.0766	$8.01 \cdot 10^{-4}$
concBigram	0.8443	0.7677	0.0766	$8.01 \cdot 10^{-4}$
concFullBigram	0.7768	0.6954	0.0814	$8.93 \cdot 10^{-4}$
conceptFull	0.809	0.7177	0.0913	$1.17 \cdot 10^{-3}$
concept + stemBigram	0.9302	0.838	0.0922	$8.39 \cdot 10^{-4}$
concept	0.8743	0.7792	0.0951	$1.13 \cdot 10^{-3}$
stem	0.9348	0.8131	0.1217	$1.36 \cdot 10^{-3}$

Table 5. Classification effectiveness across the ten features ordered by increasing max-min F-measure range (Δ).

Feature is the final variable of our analysis, and the one with the greatest impact on classification results. The use of the concFullMorph + stemBigram feature provide the highest F-measure (0.9647), while concFullBigram yields the lowest maximal F-measure (0.7768): a significant difference of 19.48%. The smallest variance was demonstrated by stemBigram ($2.02 \cdot 10^{-4}$), making it the most robust feature in our experiment; in addition this feature yielded a maximal F-measure of only 0.003% lower than the best value recorded in our experiments. The minimal F-measure yield by the stemBigram feature was also greater than the greatest F-measure values obtained when using half of the features investi-

gated in our study. These results provide strong indication that, of the variables analysed, the choice of feature provides the greatest contribution to the classification effectiveness.

5 Conclusions

Timely processing of cancer notifications is critical for timely reporting of cancer incidence and mortality. Death certificates are a rich source of data on cancer mortality. Cancer registries acquire free-text death certificates on a regular (e.g. fortnightly) basis. However, the cause of death information needs to be classified to facilitate reporting of cancer mortality. Cause of death information classified using ICD-10 codes is only available on an annual basis. In this paper we investigated the automatic classification of death certificates to individuate cancer notifiable cause of deaths. The investigated approaches achieved overall strong classification effectiveness, with a support vector machine classifier trained with token bigram features and information from the SNOMED CT medical ontology, and weighted by their frequency in the documents yielding an F-measure of 0.9647. The choice of features, rather than that of classifiers or weighting schema, was found to be the determining factor for high effectiveness.

Future efforts will be directed towards an in depth error analysis, in particular examining the distance between the prediction produced by a classifier and the decision threshold. We also plan to extend the investigation to predict the actual ICD-10 codes associated to cause of death related to cancer, so as to further assist clinical coders in processing cancer notifications.

References

1. Nguyen, A., Moore, J., Lawley, M., Hansen, D., Colquist, S.: Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. In: Health Informatics Conference. (2011) 117–124
2. Zuccon, G., Nguyen, A., Bergheim, A., Wickman, S., Grayson, N.: The impact of OCR accuracy on automated cancer classification of pathology reports. Studies in health technology and informatics **178** (2012) 250
3. D'Avolio, L., Nguyen, T., Farwell, W., Chen, Y., Fitzmeyer, F., Harris, O., Fiore, L.: Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). Journal of the American Medical Informatics Association **17**(4) (2010) 375–382
4. Harris, K.: Selected data editing procedures in an automated multiple cause of death coding system. In: Proceedings of the Conference of European Statistics. (1999)
5. Davis, K., Staes, C., Duncan, J., Igo, S., Facelli, J.: Identification of pneumonia and influenza deaths using the death certificate pipeline. BMC Medical Informatics and Decision Making **12**(1) (2012) 37
6. Nguyen, A.N., Lawley, M.J., Hansen, D.P., Bowman, R.V., Clarke, B.E., Duhig, E.E., Colquist, S.: Symbolic rule-based classification of lung cancer stages from free-text pathology reports. Journal of the American Medical Informatics Association **17**(4) (2010) 440–445
7. Witten, I., Frank, E., Hall, M.: Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques. Morgan Kaufmann (2011)

Clinician-Driven Automated Classification of Limb Fractures from Free-Text Radiology Reports

Amol Wagholikar[1], Guido Zuccon[1], Anthony Nguyen[1],
Kevin Chu[2], Shane Martin[2], Kim Lai[2], Jaimi Greenslade[2]

[1]The Australian e-Health Research Centre, Brisbane, CSIRO
{amol.wagholikar,guido.zuccon,anthony.nguyen}@csiro.au
[2] Department of Emergency Medicine, RBWH, Brisbane, Queensland Health
{kevin_chu,shane_martin}@health.qld.gov.au
{kim_lai,jaimi_greenslade}@health.qld.gov.au

Abstract. The aim of this research is to report initial experimental results and evaluation of a clinician-driven automated method that can address the issue of misdiagnosis from unstructured radiology reports. Timely diagnosis and reporting of patient symptoms in hospital emergency departments (ED) is a critical component of health services delivery. However, due to disperse information resources and vast amounts of manual processing of unstructured information, a point-of-care accurate diagnosis is often difficult. A rule-based method that considers the occurrence of clinician specified keywords related to radiological findings was developed to identify limb abnormalities, such as fractures. A dataset containing 99 narrative reports of radiological findings was sourced from a tertiary hospital. The rule-based method achieved an F-measure of 0.80 and an accuracy of 0.80. While our method achieves promising performance, a number of avenues for improvement were identified using advanced natural language processing (NLP) techniques.

Keywords: limb fractures, emergency department, radiology reports, classification, rule-based method, machine learning.

1 Introduction

The analysis of x-rays is an essential step in the diagnostic work-up of many conditions including fractures in injured Emergency Department (ED) patients. X-rays are initially interpreted by the treating ED doctor, and if necessary patients are appropriately treated. X-rays are eventually reported on by the specialist in radiology and these findings are relayed to the treating doctor in a formal written report. The ED, however, may not receive the report until after the patient was discharged home. This is not an uncommon event because the reporting did not occur in real-time. As a result, there are potential delays in the diagnosis of subtle fractures missed by the treating doctor until the receipt of the radiologist's report. The review of x-ray reports is a necessary practice to ensure fractures and other conditions identified by the radiolo-

gist were not missed by the treating doctor. The review requires the reading of the free-text report. Large "batches" of x-rays are reviewed often days after the patient's ED presentation. This is a labour intensive process which adds to the diagnostic delay. The process may be streamlined if it can be automated with clinical text processing solutions. These solutions will minimise delays in diagnosis and prevent complications arising from diagnostic errors [1-2]. This research aims to address these issues through the application of a gazetteer rule-based approach where keywords that may suggest the presence or absence of an abnormality were provided by expert ED clinicians. Rule-based methods are commonly used in Artificial Intelligence [3-5]. Studies have shown that rule-based methods can be applied for identifying clinical conditions from radiology reports such as acute cholecystitis, acute pulmonary embolism and other conditions [6]. The purpose of these methods is to simulate human reasoning for any given information processing task to achieve full or partial automation.

2 Related Work

Previous studies that focused on the problem of identification of subtle limb fractures during the diagnosis of ED patients showed that about 2.1% of all fractures were not identified during initial presentation to the Emergency Department [7]. A similar study about radiological evidence for fracture reports that 1.5% of all x-rays had abnormalities that were not identified in the Emergency Department records [8]. Further research also reported that 5% and 2% of the x-rays of the hand/fingers and ankle/foot from a pediatric Emergency Department had fractures missed by the treating ED doctor [9]. These small percentages of incidences may have significant impact on the overall patient healthcare as these missed fractures may develop into more complex conditions. Timely recognition of fractures is therefore important. There have been efforts to automatically detect fractures and other abnormalities from free-text radiology reports using support vector machine (SVM) and machine learning techniques[10-11]. Even though the results of machine learning based classifiers show high effectiveness, their applicability in clinical settings may be limited. Machine learning methods are data–driven, and as a result, if the training sample is not a representative selection of the problem domain, then the resulting model will not generalise. In addition, machine learning approaches are required to be retrained on new corpora and tasks and collating training data to build new classifier models can be a timely and labour intensive process. These issues provide the motivation for the investigation of rule-based methods which have the ability to model expert knowledge as easily implementable rules.

3 Methods

A set of 99 de-identified free-text descriptions of patient's limb x-rays reported by radiologists were extracted from a tertiary hospital's picture archiving and communication system (PACS). An ethics approval was granted by the Human

Research Ethics Committee at Queensland Health to use this data. The average length of free-text reports is about 52 words with total 930 unique words in the vocabulary. Some reports are semi-structured, with section headings such as “History”, “Clinical Details”, “Findings”, appearing in the text.

3.1 Ground Truth Development

One ED visiting medical officer and one ED Registrar were engaged as assessors to manually classify the patient findings. Findings were assigned to either one of the following two classes: (1) “Normal”, means identifying no fractures or dislocations and (2) “Abnormal”, identifying the presence of a reportable abnormality such as fracture, dislocation, displacement etc., which requires further follow-up. To gather ground truth labels about the data, an in-house annotation tool was developed. This tool allowed the assessors to manually annotate and classify the free-text reports into one of the two target categories. The two assessors initially agreed on the annotations of 77 of the 99 reports and disagreed on the remaining 22 reports. The disagreed reports were resolved and validated by a senior Staff Specialist in Emergency Medicine, who acted as a third assessor.

3.2 Rule-based classifier

A rule-based classifier was developed and implemented with rules as a set of key-words extracted from the x-ray reports assessment criteria as documented by the clinicians prior to the ground truth annotation task. The classifier was implemented to classify the text into “Normal” and “Abnormal” categories as shown in Table 1.

Table 1. Keywords used for building the rule-base.

Keywords	Suggested Classification
no + fracture	*Normal*
old + fracture	*Abnormal*
Fracture	*Abnormal*
x ray + follow up	*Abnormal*
Dislocation	*Abnormal*
FB	*Abnormal*
Osteomyelitis	*Abnormal*
Osteoly	*Abnormal*
Displacement	*Abnormal*
intraarticular extension	*Abnormal*
foreign body	*Abnormal*
articular effusion	*Abnormal*
Avulsion	*Abnormal*
septic arthritis	*Abnormal*
Subluxation	*Abnormal*
Osteotomy	*Abnormal*
Callus	*Abnormal*

4 Results and Discussion

Results obtained by our gazetteer rule-based approach on the dataset containing 99 radiology reports are reported in Table 2, along with the performance of a Naïve Bayes classifier that was used to classify on the same dataset [12]. The Naïve Bayes classifier was trained and evaluated using a 10-fold cross validation approach. This approach used 90% of reports for training and subsequently evaluated on the remaining 10% within each cross validation fold. The average of the evaluation results across the 10 folds was reported as the classifier's performance. A set of stemmed tokens in combination with high order semantic features such as SNOMED CT concepts related to morphological abnormalities and disorders generated by the Medtex system [13] were used to represent the reports. Classification results were evaluated in terms of F-measure and accuracy (see Table 2). The number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) instances were also reported.

Table 2. Classification results obtained by rule-based and NB classification

Method	F-measure	Accuracy	TP	TN	FP	FN
Rule-based	0.80	0.80	39	40	11	9
Naive Bayes	0.92	0.92	44	47	4	4

The rule-based system classified 49 reports as "Normal". Thirty-three of these were classified as "normal" due to the "no + fracture" rule. The remaining 16 reports did not match any rule, and thus were classified as "normal" (i.e. "no rule fired"). The high false negative count from the rule-based system suggests that the keywords that were used to characterise "Abnormal" cases by the clinician were not complete or adequate to capture all possible cases of abnormalities. Although the proposed keyword rule-based approach is simplistic but shows promise, advanced Natural Language Processing techniques such as those adopted in Medtex [14] can be used to improve classification performances. More keywords can also be learnt using computational linguistic methods, such as the Basilisk bootstrapping algorithm [15].

5 Conclusion and Future Research

This work has described an initial investigation of a clinician-driven rule-based method for automatic classification of free-text limb fracture x-ray findings. We described a simple keyword spotting approach where keywords were derived from classification criteria provided by clinicians. The rule-based classification method achieved promising results with F-measure performances of 0.80 and an accuracy of 0.80. As future work, the research will aim to improve the simple keyword approach with more advanced clinical text processing techniques to complement the proposed rule-based classification method. The possible integration of our method in real-life workflow of hospital emergency departments will also be considered.

Acknowledgements. The authors are thankful to Bevan Koopman for feedbacks on earlier draft of this paper. This research was supported by the Queensland Emergency Medicine Research Foundation Grant, EMPJ-11-158-Chu-Radiology.

References

1. James M. R., Bracegirdle A. and Yates D. W. X-ray reporting in accident and emergency departments – an area for improvements in efficiency. Arch Emerg Med, 8:266–270, 1991.
2. Siegel E., Groleau G., Reiner B. and Stair T. Computerized follow-up of discrepancies in image interpretation between emergency and radiology departments. J Digit Imaging, 11:18–20, 1998.
3. Long W.J, et al. Reasoning requirements for diagnosis of heart disease. Artificial Intelligence in Medicine, 10(1), pp. 5–24, 1997.
4. Harleen K., Siri Krishan W.Empirical Study on Applications of Data Mining Techniques in Healthcare, Journal of Computer Science 2 (2): 194-200, pp.1549-3636, 2006.
5. Subhash Chandra, N., Uppalaiah, B., Charles Babu, G., Naresh Kumar, K., Raja Shekar P. General Approach to Classification: Various Methods can be used to classify X-ray images, IJCSET, Vol 2, Issue 3,933-937, March 2012.
6. Lakhani P, Kim W, Langlotz CP. Automated detection of critical results in radiology reports. J Digit Imaging 25(1):30–36, 2012.
7. Cameron MG. Missed fractures in the emergency department. Emerg Med (Fremantle), 6:3, 1994.
8. Sprivulis P. and Frazer A. Same-day x-ray reporting is not needed in well supervised emergency departments. Emerg Med (Fremantle), 13:194–197, 2001.
9. Mounts J., Clingenpeel J., and E. Byers E. McGuire and Kireeva Y. Most frequently missed fractures in the emergency department. Clin Pediatr (Phila), 50:183–186, 2011.
10. De Bruijn B., Cranney A., O'Donnell S., Martin J.D. and Forster A.J. Identifying wrist fracture patients with high accuracy by automatic categorization of x-ray reports. Journal of the American Medical Informatics Association (JAMIA), 13(6):696–698, 2006.
11. Thomas B.J., Ouellette H., Halpern E.F. and Rosenthal D.I. Automated computer-assisted categorization of radiology reports. American Journal of Roentgenology, 184(2):687–690, 2005.
12. Zuccon G, Wagholikar A, Nguyen A, Chu, K, Martin S, Greenslade J., Identifying Limb Fractures from Free-Text Radiology Reports using Machine Learning, Technical Report, CSIRO, 2012.
13. Nguyen AN, Lawley MJ, Hansen DP, et al. A simple pipeline application for identifying and negating SNOMED clinical terminology in free text. Proceedings of the Health Informatics Conference; August 2009, Canberra, Australia; 188–93, 2009.
14. Nguyen A, Lawley, M., Hansen, D., Bowman, R., Clarke, B., Duhig, E., Colquist, S. Symbolic Rule-based Classification of Lung Cancer Stages from Free-Text Pathology Reports, Journal of the American Medical Informatics Association(JAMIA), vol. 17, no. 4, pp. 440-445, July/August 2010.
15. Thelen, M., Riloff, E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts, Proceedings of the ACL-02 conference on Empirical methods in natural language processing, p.214-221, July 06, 2002

Using Prediction to Improve Elective Surgery Scheduling

Zahra Shahabi Kargar[1, 2], Sankalp Khanna[1, 2], Abdul Sattar[1]

[1] Institute for Integrated and Intelligent Systems, Griffith University, Australia
{Zahra.Shahabikargar, A.Sattar@griffith.edu.au
[2] The Australian e-Health Research Centre, RBWH, Herston, Australia
{Sankalp.Khanna}@csiro.au

Abstract. Stochastic activity durations, uncertainty in the arrival process of patients, and coordination of multiple activities are some key features of surgery planning and scheduling. In this paper we provide an overview of challenges around elective surgery scheduling and propose a predictive model for elective surgery scheduling to be evaluated in a major tertiary hospital in Queensland. The proposed model employs waiting lists, peri-operative information, workload predictions, and improved procedure time estimation models, to optimise surgery scheduling. It is expected that the resulting improvement in scheduling processes will lead to more efficient use of surgical suites, higher productivity, and lower labour costs, and ultimately improve patient outcomes.

Keywords: Surgery scheduling, Predictive optimisation, Waiting list

1 Introduction

Ageing population and higher rates of chronic disease increase the demand on health services. The Australian Institute of Health and Welfare reports a 3.6% per year increase in total elective surgery admissions over the past four years [1]. These factors stress the need for efficiency and necessitate the development of adequate planning and scheduling systems in hospitals. Since operating rooms (ORs) are the hospital's largest cost and revenue centre that has a major impact on the performance of the hospital, OR scheduling has been studied by many researchers.

The surgery scheduling problem deals with the allocation of ORs under uncertain demand in a complex and dynamic hospital environment to optimise use of resources. Different techniques such as Mathematical programming[2-4], simulation [5, 6], Meta-heuristics [5, 7] and Distributed Constraint Optimization [8] have been proposed to address this problem. However most current efforts to solve this problem either make simplifying assumptions (e.g. considering only one department or type of surgery [4]), or employ theoretic data [3, 5] which make them difficult to use in hospitals.

In this paper, we propose a prediction based methodology for surgery scheduling to address the above limitations. By using predicted workload information and retrospective analysis of waiting lists and theatre utilization, we predict a theatre template representing optimal case mix. The proposed model also employs accurate estimation of procedure time and predicted workload information to drive optimal elective surgery scheduling, and help hospitals fulfil National Elective Surgery Targets (NEST) [1].

2 Elective Surgery Scheduling at the Evaluation Hospital

Long waiting lists for elective surgery in Australian hospitals during recent years has driven a nationwide research agenda to improve the planning, management and delivery of health care services. This work is to be evaluated at a major tertiary hospital which has a total of 15 operating theatres performing 124 elective operating sessions and 23 emergency sessions per week. Currently allocation of available elective operating sessions at the hospital have been broken down to different specialties and teams of surgeons based on a static case mix planning. This static allocation of available sessions between emergency and elective patients and among different departments results in underutilization or cancellation due to demand fluctuations. Also, the allocation of patients to theatres is carried out without considering the uncertainty and possible changes that might happen. Procedure times are estimated by using generic data or recommended by relevant surgeons not based on individual patient and surgery characteristics. Patients are booked into schedules in a joint process between surgeons and the booking department. Due to the dynamic environment and rapid changes, these schedules need to be updated quickly. Usually department managers have regular meetings to make any changes needed. Department managers try to locally optimise their department goals, but since there is no global objective usually these solutions are not the optimal global solutions.

3 An Optimal Surgery Scheduling Model

Although the surgery scheduling problem has been well addressed in literature, it still remains an open problem in Operations Research and Artificial Intelligence. Despite the dynamic nature of the hospital environment, the majority of previous studies ignore the underlying uncertainty. This leads to simplistic models that are not applicable in real world situations.

3.1 Current State of the Art

Cardoen et al. present a comprehensive literature review on operating room scheduling including different features such as performance measures, patient classes, solution technique and uncertainty [9]. One of the major issues associated with the development of accurate operating room schedules or capacity planning strategies is the uncertainty inherent to surgical services. Uncertainty and variability of frequency and distribution of patient arrivals, patient conditions, and procedure durations, as well as ‘‘add-on’’ cases are some instances of uncertainty in surgery scheduling [10]. Among them stochastic arrival and procedure duration are two type of uncertainty studied by many researchers. Procedure duration depends on several factors such as experience of the surgeon, supporting staff, type of anaesthesia, and precondition of the patient. Devi et al. estimate surgery times by using Adaptive Nero Fuzzy Inference Systems, Artificial Neural Networks and Multiple Linear Regression Analysis [2] but they just focus on one department and use a very limited sample to build and validate their model. Lamiri et al. developed a stochastic model for planning elective surgeries under uncertain demand for emergency surgery [3]. Lamiri et al. also address the elective surgery planning under uncertainties related to surgery times and emergency surgery demands by combining Monte Carlo simulation and a column generation approach[5]. Although their method addresses uncertainties, it is based on theoretic data and it has not been tested on real data. What is needed is a whole of theatre approach to provide better prediction of surgery time, incorporation of predicted workload in planning the weekly surgery template, and target guided optimization to ensure optimal allocation of resources.

3.2 Proposed Method

To improve the planning and optimization tasks underlying the process, we propose a two stage methodology for elective surgery scheduling. As a first stage, predicted workload information (drawn from Patient Admission Prediction Tool [11] currently used at the evaluation hospital), current Waiting List information and Historic utilization information is used to manage theatre allocation and case mix distribution for each week (see Figure 1). This allows the prediction based sharing of theatres between elective and emergency surgery, and allocation of theatre time to surgery teams/departments and results in a theatre schedule template that works better than a static allocation model (as demonstrated by Khanna et al. [8]).

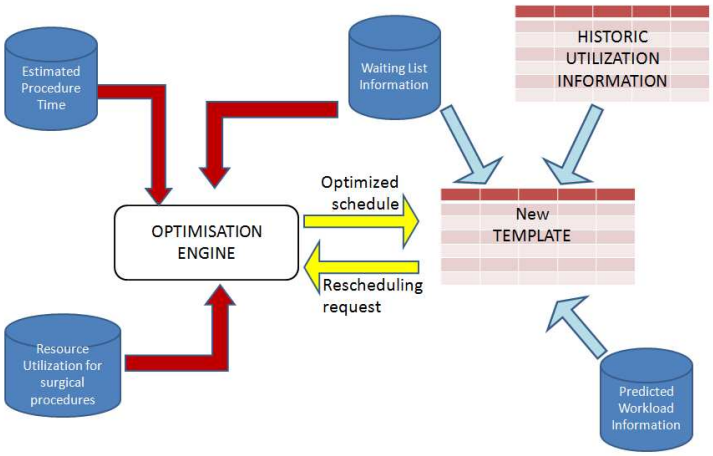


Figure 1. Proposed Methodology for Improving Surgery Scheduling

In the second stage of the process, the allocation of patients to the weekly theatre schedule is guided by an improved prediction algorithm to estimate the surgery duration. The algorithm takes into account current patient, surgery, and surgeon information and related historic peri-operative information to forecast the planned procedure time. Incorporating NEST compliance in the optimization function and improved resource estimation deliver further improvements to the scheduling process and help deliver a more robust and optimal schedule (Figure 1). We are currently working towards collecting over 5 years of surgery scheduling, waiting list and peri-operative information for the evaluation hospital from the corporate information systems. This data will be used for modelling and independently validating the prediction algorithms and building historic resource utilization knowledge banks to guide other stage of the scheduling process.

4 Conclusion

The proposed model has the potential to improve elective surgery scheduling by providing more accurate procedure time estimation and predicting arrival demand of elective and emergency patients.

References

1. Health, D.o., *Expert Panel Review of Elective Surgery and Emergency Access Targets Under the National Partnership Agreement on Improving Public Hospital Services.* 2011.
2. Devi, S.P., K.S. Rao, and S.S. Sangeetha, *Prediction of surgery times and scheduling of operation theaters in ophthalmology department.* J Med Syst, 2012. **36**(2): p. 415-30.
3. Lamiri, M., Xiaolan Xie, and Shuguang Zhang, *Column Generation Approach to Operating Theater Planning with Elective and Emergency Patients.* IIE Transactions, 2008. **40**(9): p. 838–852.
4. Pérez Gladish, B., et al., *Management of surgical waiting lists through a Possibilistic Linear Multiobjective Programming problem.* Applied Mathematics and Computation, 2005. **167**(1): p. 477-495.
5. Lamiri, M., J. Dreo, and Xiaolan Xie. *Operating Room Planning with Random Surgery Times*. in *IEEE International Conference On Automation Science and Engineering*. 2007. Scottsdale, AZ, USA.
6. S.M. Ballard, M.E.K. *The use of simulation to determine maximum capacity in the surgical suite operating room*. in *Proceedings of the 2006 Winter Simulation Conference*. 2006.
7. Fei, H., Nadine Meskens, and Chengbin Chu. *An Operating Theatre Planning and Scheduling Problem in the Case of a 'Block Scheduling' Strategy*. in *International Conference on Service Systems and Service Management*. 2006.
8. Khanna, S., Abdul Sattar, Justin Boyle, David Hansen, and Bela Stantic. *An Intelligent Approach to Surgery Scheduling*. in *Proceedings of the 13th International Conference on Principles and Practice of Multi-Agent Systems*. 2012. Berlin.
9. Cardoen, B., Erik Demeulemeester, and Jeroen Beliën, *Operating Room Planning and Scheduling: A Literature Review.* European Journal of Operational Research, 2010. **201**(3): p. 921–932.
10. May, J.H., William E. Spangler, David P. Strum, and Luis G. Vargas., *The Surgical Scheduling Problem: Current Research and Future Opportunities.* Production and Operations Management, 2011. **20**(3): p. 392–405.
11. Boyle, J., M. Jessup, J. Crilly, D. Green, J. Lind, M. Wallis, P. Miller, and G. Fitzgerald, *Predicting Emergency Department Admissions.* Emergency Medicine Journal 2011. **29**(5): p. 358–365.

If you fire together, you wire together

Prajni Sadananda[1], Ramakoti Sadananda[2, 3]

[1] Department of Anatomy and Neuroscience, University of Melbourne Australian, Australia
prajni.sadananda@unimelb.edu.au
[2] Institute for Integrated and Intelligent Systems, Griffith University, Australia
[3] NICTA, Sydney Australia
rsadananda@griffith.edu.au

The intention of this paper is to stimulate discussion on Hebb's Law and its pedagogic implications.

At a basic cellular level, Hebb's Law states that is Cell A and Cell B persistently fire, the connection between them strengthens. Figure 1 illustrates the interactions. This is a cellular levels process, suggesting that brain processes that occur repeartedly tend to become grafted together [1].

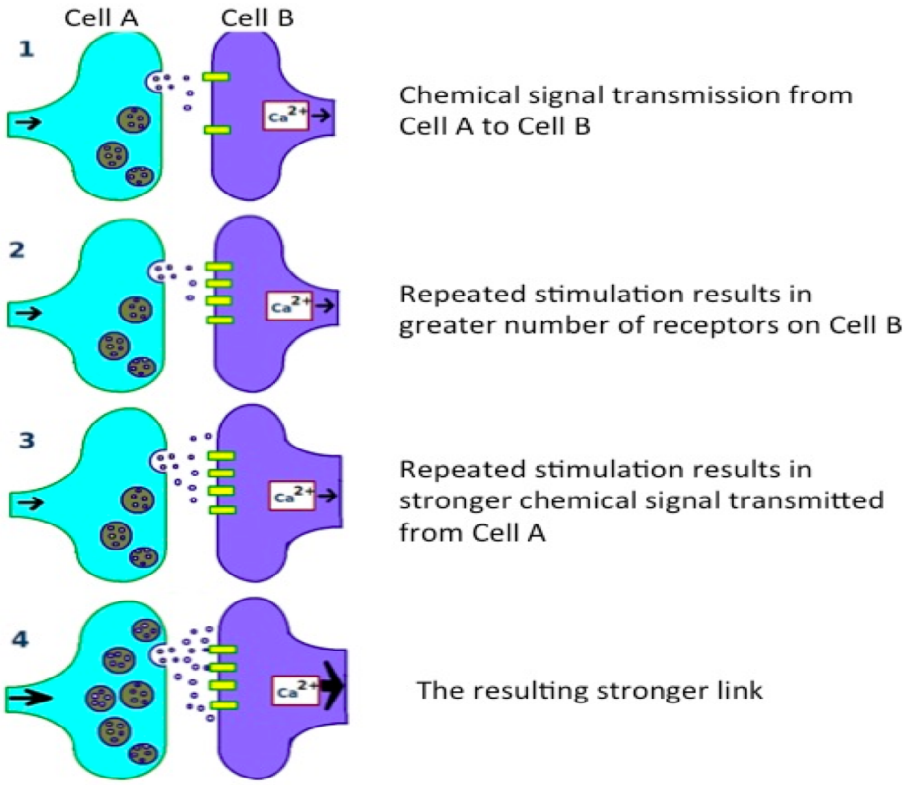


Fig 1. Hebb's Law. Repeated stimulation results in a stronger signal

This scientific theory explains the adaptation of neurons during the learning process. Importantly, this type of plasticity does not involve increasing the number of cells, but rather strengthening the existing cells' connectivity. Understanding such biological phenomena opens up new paradigms and laws that AI can utilise. The question is whether Hebb's law would stand at a higher level of abstraction? There are suggestive, but not conclusive indications. For example, a friendship is considered stronger with time, indicating a strengthening of wiring between the friends.

Models based on "firing together to wire together" have been suggested in health and therapy [2]. For example, if a patient presents with a mental trauma that causes extreme anger, the therapist introduces a counter and positive stimulus that occurs whenever the anger occurs. Both (anger and the positive stimulus) are repeated over and over again, thus following Hebb's Law and adding strength to this connection between the two stimuli, resulting in relief to the patient.

This also implies causal and temporal conjectures based on causality. The causality is in the firing sequence; that if A fires first and then B fires, A is the cause. If B fires before A, a reverse interpretation is possible that may decrease the strength between them. There seems evidence to suggest that the "firing" and "wiring" may be a sequential process.

Causality is a subject of intense philosophical interest from ancient times. Most causal models are rule-based systems. They demand descriptions of the world at two points in time – a before and an after. Two problems arise here: the practical computational compulsions make these rules crudely simplistic. In addition, it is challenging to incorporate temporal effects within the framework of rule based systems. Hebb's law, while suggesting causality, does not provide any quantification. Thus, it is unlikely that an alternative formulation of causation would emerge from Hebb's law alone. We may look for another, additional neural network perspective of causation here.

Nevertheless, causality as implied with Hebb's law has been used in scientific research and therapeutics to a large' extent. For example, oftentimes doctors complain of their patients being unable to add minor and incremental changes in their daily routines (such as exercise). Understanding Hebb's law will open new insights into why this might be so. It is possible that the patient is not yet "wired" in this activity and requires more "firing" before

these changes can be established. An avenue for AI research is to aide in the development of tools to help such people to “re-wire”.

Indeed, such tools exist to some extent to treat spinal cord injured patients who have lost motor control of their limbs. In a non-injured situation, the brain delivers pulses to the lower limbs in a rhythmic/patterned fashion to allow walking action. Once a spinal injury occurs, the connectivity from the brain to the limbs is lost, thereby leaving the patient immobile. Stimulators are often placed below the level of the injury, which deliver patterned pulses in a similar manner to what the brain was previously doing. Over a period of time, a spinal pattern generator emerges, which thus allows some motion of the lower limbs [3]. This area of research is as yet in its infancy and calls for a better, more intelligent systems to aide these patients.

Conclusions:

Artificial intelligence in health opens up chapters of great opportunities and exciting challenges. The logical calculus articulated by McCulloch and Pitts [4] forms the initial basis for both Symbolic and Connectionist AI. Since then a number of paradigms have emerged on all aspects of AI and relating to health and health care. The emergence of the convergence of computing and communication provides us boundless opportunities to exploit these paradigms and discover the new ones.

Referenceă:

1. Hebb, D. O.: Organization of Behavior: a Neuropsychological Theory. John Wiley, New York (1949).
2. Atkinson, B., Atkinson, L., Kutz, P., Lata, L., Lata, K.W., Szekely, J., Weiss, P.: Rewiring Neural States in Couples Therapy: Advances from Affective Neuroscience. In: Journal of Systemic Therapies. 24, 3-13 (2005)
3. Edgerton, V.R., Roy, R.R.: A new age for rehabilitation. Eur J Phys Rehabil Med. 48, 99-109 (2012)
4. McCulloch, W.S., Pitts, W.: A logical Calculus of the ideas immanent in nervous activity, Bulletin of mathematical Biophysics. 5, 115-137 (1943).

NOTES

NOTES