University of South Florida

Digital Commons @ University of South Florida

USF Tampa Graduate Theses and Dissertations

USF Graduate Theses and Dissertations

January 2013

Classification Models in Clinical Decision Making

Eleazar Gil-Herrera
University of South Florida, eleazar@mail.usf.edu

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Part of the Computer Engineering Commons, Industrial Engineering Commons, and the Medicine and Health Sciences Commons

Scholar Commons Citation
Gil-Herrera, Eleazar, "Classification Models in Clinical Decision Making" (2013). *USF Tampa Graduate Theses and Dissertations.*
https://digitalcommons.usf.edu/etd/4895

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.

Classification Models in Clinical Decision Making

by

Eleazar Gil-Herrera

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Industrial and Management Systems Engineering
College of Engineering
University of South Florida

Major Professor: Ali Yalcin, Ph.D.
Laura E. Barnes, Ph.D.
Grisselle Centeno, Ph.D.
Benjamin Djulvegovic, Ph.D.
Peter Fabri, Ph.D.
Autar Kaw, Ph.D.
Athanasios Tsalatsanis, Ph.D.

Date of Approval:
November 14, 2013

Keywords: Patient-centered prognostic models, Clinical data analysis, Rough set theory, Hospice referral, Knowledge discovery

Copyright © 2013, Eleazar Gil-Herrera

Dedication

To my wife, my love, best friend and complement of my life and to my two little angels who put joy and motivation in this work. This achievement is also yours.

To the Almighty, guidance and support of my life, because everything in His hands becomes good.

Acknowledgments

I am deeply grateful to my advisor, Dr. Ali Yalcin, for his unconditional guidelines and motivation during my doctoral studies. I thank him specially for his patience, time and thoroughness in reviewing my writings. Thank you for your dedication to make me a better researcher and professional. I am also in debt with the faculty and staff of the Clinical and Translational Science Institute at University of South Florida, special thanks to Dr. Djuvelgovic, Sakis, and Laura, for their support and consideration.

My thanks also goes to all the friends I met on this warm place. Thank you guys for your prayers, blessings, good wishes and especially for making this time more bearable.

Table of Contents

Abstract

In this dissertation, we present a collection of manuscripts describing the development of prognostic models designed to assist clinical decision making. This work is motivated by limitations of commonly used techniques to produce accessible prognostic models with easily interpretable and clinically credible results. Such limitations hinder prognostic model widespread utilization in medical practice.

Our methodology is based on Rough Set Theory (RST) as a mathematical tool for clinical data analysis. We focus on developing rule-based prognostic models for end-of life care decision making in an effort to improve the hospice referral process. The development of the prognostic models is demonstrated using a retrospective data set of 9,103 terminally ill patients containing physiological characteristics, diagnostic information and neurological function values.

We develop four RST-based prognostic models and compare them with commonly used classification techniques including logistic regression, support vector machines, random forest and decision trees in terms of characteristics related to clinical credibility such as accessibility and accuracy. RST based models show comparable accuracy with other methodologies while providing accessible models with a structure that facilitates clinical interpretation. They offer both more insight into the model process and more opportunity for the model to incorporate personal information of those making and being affected by the decision.

Chapter 1: Introduction

1.1 Clinical Decision Making and Requirements for Prognostic Models

The Institute of Medicine (IOM) emphasizes the customization of health care to be responsive to individual patient preferences, needs, and values [1]. Therefore, treatment recommendations and decision-making are based in response to individual patient indicators of health state. This vision of personalized health care requires new methodologies for developing patient-centered prognostic and diagnostic models resulting in the selection of appropriate treatment for each patient.

To be accepted by physicians and patients and to be used in practice, prognostic and diagnostic models must have clinical credibility [2]. That is, in addition to accurate prognostication, a model should be traceable in its structure, allowing complete insight to the prognostic process; the variables in the model should possess clinical relevance and its results should be interpretable thus facilitating explanation of the prognosis.

1.2 Prognostic Models in Medicine: Strengths and Weaknesses of Widely Used Methods

Prognostic and diagnostic models assist physicians in making more accurate predictions and are shown to be superior to physicians' prognostication alone [3]. In addition, the accuracy of the models is further improved when combined with physicians' estimates [4–6]. Widely used models based on statistical approaches make assumptions regarding the relationship between the prognostic factors and the outcome variable. When these assumptions are violated, the resultant model is no longer representative of the data. As an example, logistic regression assumes a linear relationship existing between a given prognostic factor and the logit form of the outcome variable [7]. If the relationship is not linear, the statistical significance of the logistic regression coefficient related to that prognostic factor may be inaccurate [8]. Artificial intelligence approaches, such as neural networks and support vector machines are designed to cope with complex predictor-outcome variable relationship and are shown to be efficient in managing large amounts of information. However, as black-box methods, they offer little insight into the process of prediction and are difficult to interpret. None of these methods provide traceable and accessible results, which lead to models that lack credibility.

1.3 Characteristics of Datasets Representing Clinical Information

Hood et al.[9], estimate that in 10 years, a virtual cloud of billions of data points including information about genome sequence, images, demography, diagnostic tests and environmental data will represent a patient's medical record. The collection of such records will constitute a clinical dataset. Such a big and heterogeneous clinical data is prone to noise and present inconsistencies resulting from the inherent complex reality of illness and human physiology.

The complexity of clinical data due to its volume and heterogeneity causes the data to lack a canonical form [10]. Furthermore, the underlying conceptual structures of medicine are not easily formalized mathematically, as the medical field lacks the necessary constraints for the mathematical characterizations common to the physical sciences. As a result, many medical concepts are vaguely defined [11]. These particular characteristics of clinical data must be addressed when building prognostic models.

One of the grand challenges of personalized medicine is to reduce the dimensionality of clinical datasets and express the information in simple hypothesis about health and disease. Thus, there is a need for new mathematical methodologies to analyze heterogeneous, noisy, and inconsistent clinical data; extracting at the same time relevant information that provides insights to the diagnosis and prognosis of a disease.

1.4 Problem Description

Disease diagnosis and prognosis can be seen as a classification process with a discrete outcome variable d representing the result obtained for a given patient. In the case of diagnosis, the binary outcome $d = 1$ represents a patient with a positive result for a given disease. In prognostic models, $d = 1$ denotes the occurrence of an event in a patient within a certain follow up period; for example, recurrence of a disease, re-operation or death.

Each patient record in a clinical dataset can be represented by a tuple (x, d), where x represents the set of characteristics that describe a patient. The objective of a prognostic model is to estimate the relationship between x and d, and then use this information to predict the value of d given the values of x corresponding to new patients.

Current methodologies for developing prognostic models are focused at the patient population level, where a unique model characterizes the entire population. In contrast, the new trends of personalized health care require mathematical models to make predictions considering individual patients' characteristics, and as required, make optimal decisions tailored for each patient.

1.5 Goals and Objectives

The goal of this dissertation is to design and develop classification models consistent with the current trends for improving health care. That is, patient-centered classification models with features that allow the model to be clinically credible and useful in clinical practice.

To achieve this goal, we defined the following objectives:

1. Evaluate different classification methodologies with respect to their accuracy and accessibility, considering clinical datasets that exhibit inconsistencies and complex predictor-outcome relationships.
2. Develop accessible and accurate classification models for non-trivial clinical datasets.
3. Design and develop a methodology for analyzing clinical datasets at the individual patient level and develop patient-centered classification models.

1.6 Developing Classification Models for Hospice Referral

We focus our dissertation in the development of patient-centered classification models in an effort to improve the hospice referral process. Hospice is designed to provide quality of life and support for terminally ill patients and their families. In the U.S., Medicare regulations stated that a patient should be referred to hospice if his/her life expectancy is less than 6 months as certified by the primary physician.

Despite the well-documented advantages of hospice services, terminally ill patients do not reap the maximum benefits of hospice care with the majority of them being referred to hospice either prematurely or too late. A premature hospice referral is translated to patients losing the opportunity to receive potentially effective treatment, which may have prolonged their lives. Conversely, late hospice referral reduces the quality of life for patients and their families. It is apparent that accurate prognostication of life expectancy is of vital importance for all parties involved in the hospice referral process (e.g. patients, their families, and their physicians).

In this work, rather than predicting life expectancy, we want to determine whether the death event occurs before the six month period to consider a patient as a hospice candidate. We define the binary variable d to represent the event of the death of an individual patient, where, $d = 1$ represents a patient who does not survive the period of six months.

1.7 Summary of the Manuscripts

The manuscripts in this dissertation present in detail the development of different Rough Set Theory based classification models. Below is a summary of the manuscripts' contents describing at the same time the progression of our research towards a patient-centered and clinically credible classification models. The complete versions of these manuscripts are in the Appendix section.

In the first manuscript, *Predicting Academic Performance Using a Rough Set Theory-Based Knowledge Discovery Methodology* (Appendix B), RST is used to predict student performance in an engineering course. This initial exercise, demonstrates the strengths of the RST approach to analyze datasets and develop classification models that represents the characteristics of individuals. We are able to extract decision rules with minimal information that classify new students as being successful or unsuccessful in the class with notable classification accuracy. In addition, the results of the model in the form of if-then decision rules provide effective decision support towards the improvement of student performance.

In medicine, applications of RST are mainly focused on the diagnosis and prognostication of diseases, where it has been demonstrated that RST is useful for extracting medical prognostic rules using minimal information. In the next four manuscripts (Appendix C-F) we focus on the development of clinically credible prognostic classification models for hospice referral. We utilize retrospective data from 9,103 terminally ill patients to demonstrate the design and implementation of a classifier based on RST to determine potential candidates for hospice referral.

The second manuscript, *Rough Set Theory Based Prognostication of Life Expectancy for Terminally Ill Patients* (Appendix C), explores methodologies based on genetic algorithms and dynamic reducts for developing RST-based classification models. A unique feature of the proposed model is a condition attribute intended to represent the physicians' life expectancy estimate. By including this feature, we increase the performance of the classifier with an accuracy exceeding that of the baseline, gold standard, life expectancy prognostic model [6]. However, around 30% of the test cases, considered as new patients, remain unclassified. Having decision rules with high number of attributes, and attributes with numerous categorical levels cause the decision rules to be too specific for the training set and consequently reducing the ability of the model to classify new cases.

To address this issue, in the manuscript, entitled as: *Towards a Classification Model to Identify Hospice Candidates in Terminally Ill Patients* (Appendix D), we explore the *object related reducts* (ORR) as a method for reducing the dimensionality of the dataset. Decision rules generated by this scheme contain fewer attributes and are better suited for classifying new cases. The classification model covers 100% of the test cases and improves the overall performance. A distinctive feature of this approach is that it reveals redundancy in the condition attributes applicable for certain groups of patients. For example, we found that for some patients, the use of the *Pafi* test (blood gases) does not improve the accuracy of the prognostication. By analyzing the information in the ORRs we can identify groups of patients for whom it is possible to evade costly, invasive or unnecessary tests. One limitation of this approach is that the number of ORR and the decision rules generated can be extremely large as they depend on the number of condition attributes and its categories. This limitation reduces the model's accessibility and interpretability especially when applied to clinical datasets that typically contain large numbers of condition attributes.

In the fourth manuscript, *Rough Set Theory Based Prognostic Models for Hospice Referral* (Appendix E), we explore and evaluate the application of the classical and dominance-based RST (DRSA) to develop clinical prognostic models for hospice referral. The DRSA approach considers patients characteristics with preference-ordered values in relation to the patients' risk of death. In this case, for both the classical and the DRSA, the dimensionality reduction process is omitted as the decision rules are induced directly from the dataset. In addition, we relax the strictness of the dominance principle to induce more general rules, for each rule, the proportion information consistent with the dominance principle. Selecting an appropriate consistency level improves the model accuracy and reduces the number of unclassified patients.

The overall performance of the RST-based classifiers is compared to widely used classification approaches such as Logistic Regression, Support Vector Machines, C4.5 and Random Forrest. The results show that RST based methods perform comparable to the rest of the classification methods, while providing significant advantages in terms of traceability of the model and interpretability of the results. In particular, the DRSA method provides a set of compact, easily explainable rules that support the estimated life-expectancy classification. Inducing general rules (rules with few condition attributes) prevents overfitting the training set and results in models that are more useful in classifying new cases. However, using rules with few attributes, may cause skepticism, as some factors considered important in clinical practice may be omitted [12, 13]. Moreover, shorter rules lose the property to capture individual patient's characteristics necessary to develop a patient-centered model.

In the last manuscript, *Towards a Patient-Centered Classification Model for Hospice Referral* (Appendix F), we develop a methodology to build a patient-centered classification model. The methodology considers relevant characteristics of patients that differentiate them from the rest of patients having a different outcome. Given this information, the population of patients is divided in subgroups having similar characteristics pertaining to each group. The subgroups obtained reveal insights about the information requirements for classification of new cases. The performance of the proposed patient-centered classification model is compared with widely used classification methodologies, in terms of its accuracy, coverage and accessibility.

Chapter 2: Conclusions

Most relevant research associated with the development of prognostic models evaluates the model performance in terms of its accuracy and discrimination ability. This work, in addition, evaluates whether a prognostic model is accessible and therefore potentially useful in clinical practice. Our results demonstrate that the if-then decision rule structure offers significant advantages by increasing the accessibility of the model as the prognosis is performed using a list of readily interpretable decision rules facilitating the traceability of the results without compromising its accuracy.

In our proposed models, classification of new patients is based on a minimum set of condition attributes leading to two distinct advantages. First, it is possible to identify potentially unnecessary, expensive and/or invasive procedures that may not be necessary for classification. Second, the decision rules can be used to classify new patients even when values for some attributes are missing. This is in contrast to a logistic or Cox regression model, where complete information on all attributes is required to determine the patient prognosis.

We introduce Dominance-based Object Related Reducts (DORR) as a method to decompose a dataset into subgroups and build localized classification models. Compared to the VC-DOMLEM algorithm used for hospice referral in [14], the DORR method shows no significant improvement in the accuracy or the accessibility of the model.

However, the DORR-based subgroups provide useful information for sequential decision-making, where the objective is to determine the most appropriate strategy that maximizes the benefits for a particular patient. For example, in a disease diagnosis process, it is valuable for a physician to identify which set of tests is the most appropriate alternative for a particular patient. After performing a diagnosis test, the physician must decide whether to treat the patient immediately or continue testing.

The DORR method allows determining subgroups of patients for whom the completion of a set of diagnostic tests is indispensable for an accurate diagnosis and on the other hand patients for whom particular test results are unnecessary. Requiring one or more diagnostic tests implies migrating patients to a different subgroup, where it is possible to evaluate if acquiring more information improves the diagnosis accuracy and is beneficial for the patients. The relationship among DORR-based subgroups represents therefore paths for a sequential decision-making network where the interest is to arrive at an appropriate diagnosis accuracy level without necessarily performing the full set of tests.

Chapter 3: Limitations and Future Research

The performance of classification models is still a major issue for the targeted domain of life expectancy prognostication. Classifier performance, measured by AUC, is still sub-optimal, indicating a challenging problem in need of further research.

One area that needs to be explored is the appropriate weighting of the condition attributes in terms of their impact on the decision variable. The baseline case assumes that all the variables considered in the model are weighed equally. We believe that a careful weighting of the attributes by consulting an expert may greatly improve the classification accuracy of the models.

Another important limitation of this study is that patient-specific disease progression over time is not considered, in part due to the static nature of the data set used. Future research must address the temporal aspect of disease progression, a consideration often missing in other prognostic models for hospice referral. The progression of a terminal illness is often highly non-linear by nature and generally does not present as a steady decline over time but rather as periods of relative stability marked by turning points of acute decline. A prognostic model that takes into account this temporal aspect may possibly provide both more accurate life expectancy prognoses and more useful information for end-of-life decisions.

The DORR methodology is promising for sequential clinical decision-making as the paths defined by the subgroups relationships provide important information to construct a cost-preference network for diagnosis. Including information about patients' needs, preferences and diagnosis tests costs is valuable for obtaining a patient-centered diagnosis strategy by optimizing the cost-preference network.

Regardless of the accuracy of any classifier, medical decisions must consider the individual patient preferences towards alternative forms of treatments. Our intent for future research is to incorporate our methodology into a patient-centered decision support system that facilitate the hospice referral process.

Finally, future work should evaluate the accessibility of decision rules in clinical practice through testing the model by practitioner clinicians.

References

[1] I. of Medicine, Media Reviews, Journal for Healthcare Quality 24 (5) (2002) 52–54, ISSN 1945-1474.

[2] J. C. Wyatt, D. G. Altman, Commentary: Prognostic models: clinically useful or quickly forgotten?, BMJ 311 (7019) (1995) 1539–1541.

[3] R. Dawes, D. Faust, P. Meehl, Clinical versus actuarial judgment, Science 243 (4899) (1989) 1668–1674.

[4] E. Gil-Herrera, A. Yalcin, A. Tsalatsanis, L. Barnes, D. B, Rough Set Theory based Prognostication of Life Expectancy for Terminally Ill Patients, in: IEEE Eng Med Biol Soc, 6438–6441, 2011.

[5] K. L. Lee, D. B. Pryor, F. E. Harrell, et al., Predicting outcome in coronary disease statistical models versus expert clinicians, The American journal of medicine 80 (4) (1986) 553–560.

[6] W. A. Knaus, F. E. Harrell, J. Lynn, Goldman, et al., The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults, Annals of Internal Medicine 122 (3) (1995) 191–203.

[7] M. P. LaValley, Logistic Regression, Circulation 117 (18) (2008) 2395–2399.

[8] M. Scott W., Logistic Regression Diagnostics and Problems of Inference, SAGE Publications, Inc., 2010.

[9] L. Hood, R. Balling., Revolutionizing medicine in the 21st century through systems approaches, Biotechnol J 7 (8) (2012) 992–1001.

[10] K. J. Cios, G. W. Moore, Uniqueness of medical data mining, Artificial Intelligence In Medicine 26 (1–2).

[11] P. Simons, VAGUENESS - WILLIAMSON,T, International Journal of Philosophical Studies 4 (2) (1996) 321–327.

[12] E. Steyerberg, Clinical Usefulness, in: Clinical Prediction Models, Statistics for Biology and Health, Springer New York, ISBN 978-0-387-77243-1, 281–297, 2009.

[13] M. Ebell, AHRQ White Paper: Use of clinical decision rules for point-of-care decision support, Med Decis Making 30 (6) (2010) 712–21.

[14] E. Gil-Herrera, G. Aden-Buie, A. Yalcin, et al., Rough Set Theory based Prognostic Model for Hospice Referral, Artificial Intelligence of Medicine (in second round review) .

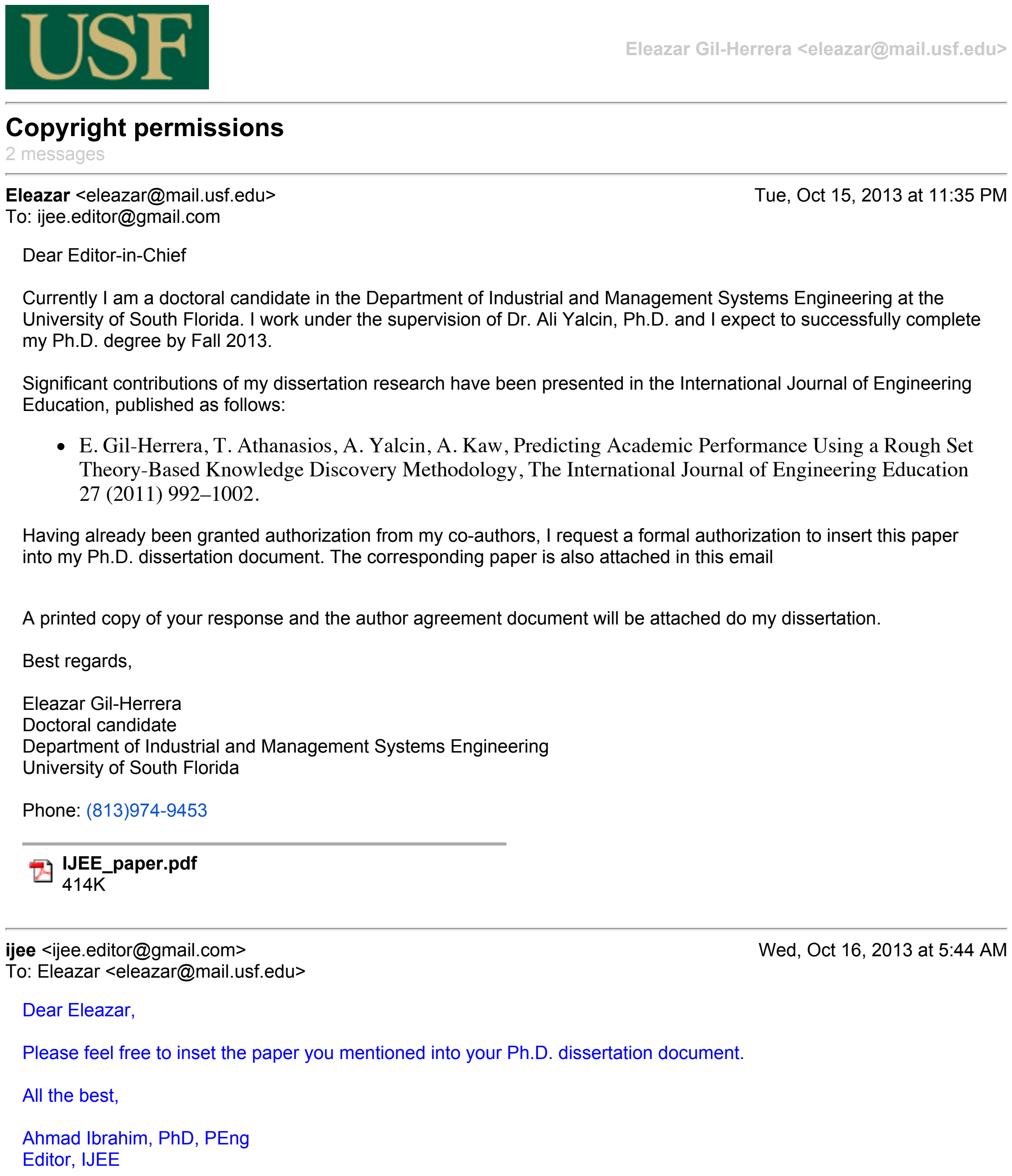
Appendices

Appendix A: Copyright Approvals

This appendix presents the written authorizations to include the following previously-published papers in this dissertation:

- *Predicting Academic Performance Using a Rough Set Theory-Based Knowledge Discovery Methodology*, published in the International Journal of Engineering Education. vol. 27, No 5, pages 992-1002, 2011 (Appendix B).
- *Rough Set Theory Based Prognostication of Life Expectancy for Terminally Ill Patients*, published in Annual International Conference Proceeding, Engineering in Medicine and Biology Society, EMBS-2011 (Appendix C).
- *Towards a Classification Model to Identify Hospice Candidates in Terminally Ill Patients*, published in Annual International Conference Proceeding, Engineering in Medicine and Biology Society, EMBS-2012 (Appendix D).

Appendix A (continued)



Eleazar Gil-Herrera <eleazar@mail.usf.edu>

Copyright permissions

2 messages

Eleazar <eleazar@mail.usf.edu> Tue, Oct 15, 2013 at 11:35 PM
To: ijee.editor@gmail.com

Dear Editor-in-Chief

Currently I am a doctoral candidate in the Department of Industrial and Management Systems Engineering at the University of South Florida. I work under the supervision of Dr. Ali Yalcin, Ph.D. and I expect to successfully complete my Ph.D. degree by Fall 2013.

Significant contributions of my dissertation research have been presented in the International Journal of Engineering Education, published as follows:

- E. Gil-Herrera, T. Athanasios, A. Yalcin, A. Kaw, Predicting Academic Performance Using a Rough Set Theory-Based Knowledge Discovery Methodology, The International Journal of Engineering Education 27 (2011) 992–1002.

Having already been granted authorization from my co-authors, I request a formal authorization to insert this paper into my Ph.D. dissertation document. The corresponding paper is also attached in this email

A printed copy of your response and the author agreement document will be attached do my dissertation.

Best regards,

Eleazar Gil-Herrera
Doctoral candidate
Department of Industrial and Management Systems Engineering
University of South Florida

Phone: (813)974-9453

IJEE_paper.pdf
414K

ijee <ijee.editor@gmail.com> Wed, Oct 16, 2013 at 5:44 AM
To: Eleazar <eleazar@mail.usf.edu>

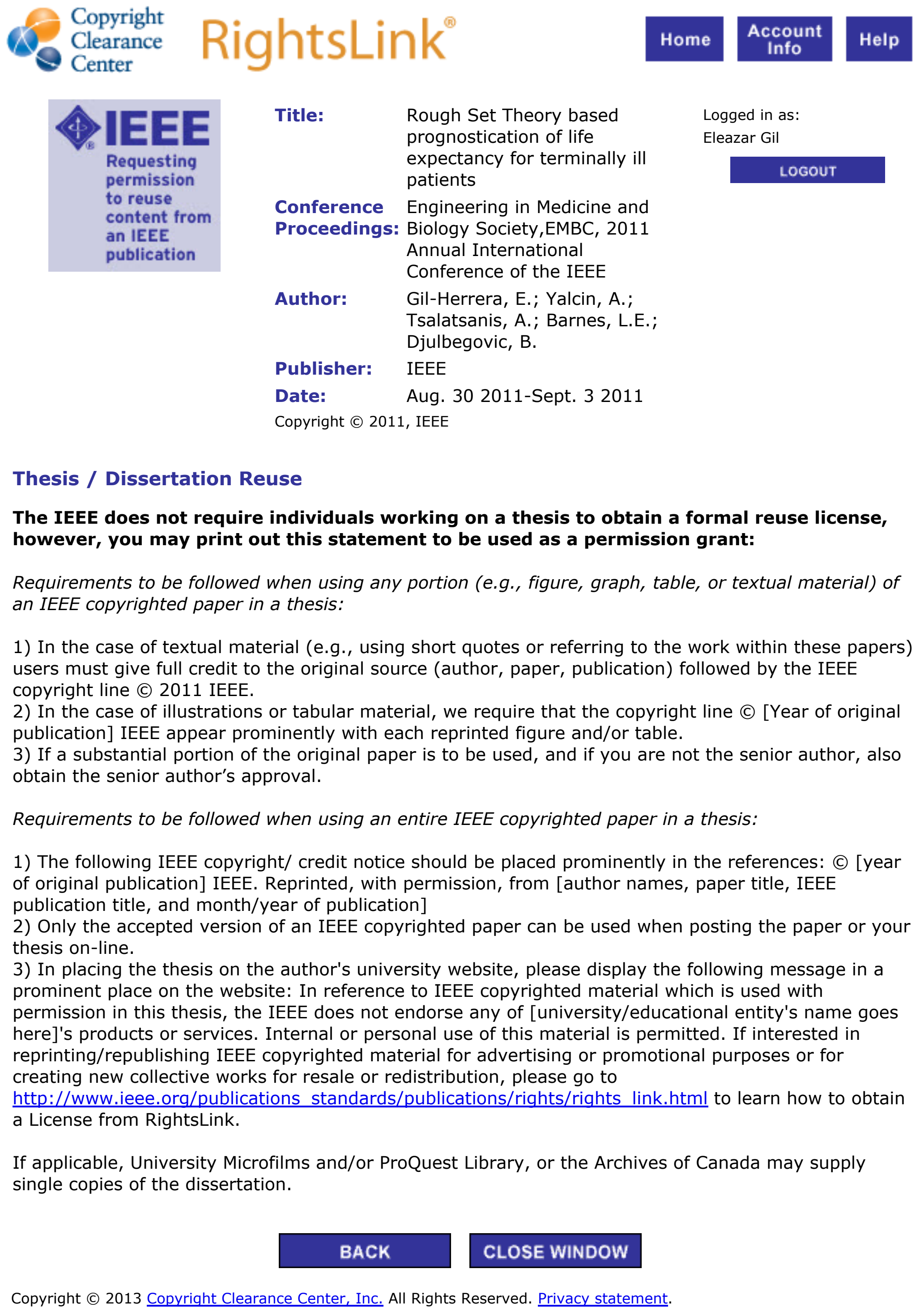
Dear Eleazar,

Please feel free to inset the paper you mentioned into your Ph.D. dissertation document.

All the best,

Ahmad Ibrahim, PhD, PEng
Editor, IJEE

Appendix A (continued)


Copyright Clearance Center
RightsLink®
Home
Account Info
Help
IEEE
Requesting permission to reuse content from an IEEE publication
LOGOUT
BACK
CLOSE WINDOW

Title: Rough Set Theory based prognostication of life expectancy for terminally ill patients

Conference Proceedings: Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE

Author: Gil-Herrera, E.; Yalcin, A.; Tsalatsanis, A.; Barnes, L.E.; Djulbegovic, B.

Publisher: IEEE

Date: Aug. 30 2011-Sept. 3 2011

Copyright © 2011, IEEE

Logged in as:
Eleazar Gil

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

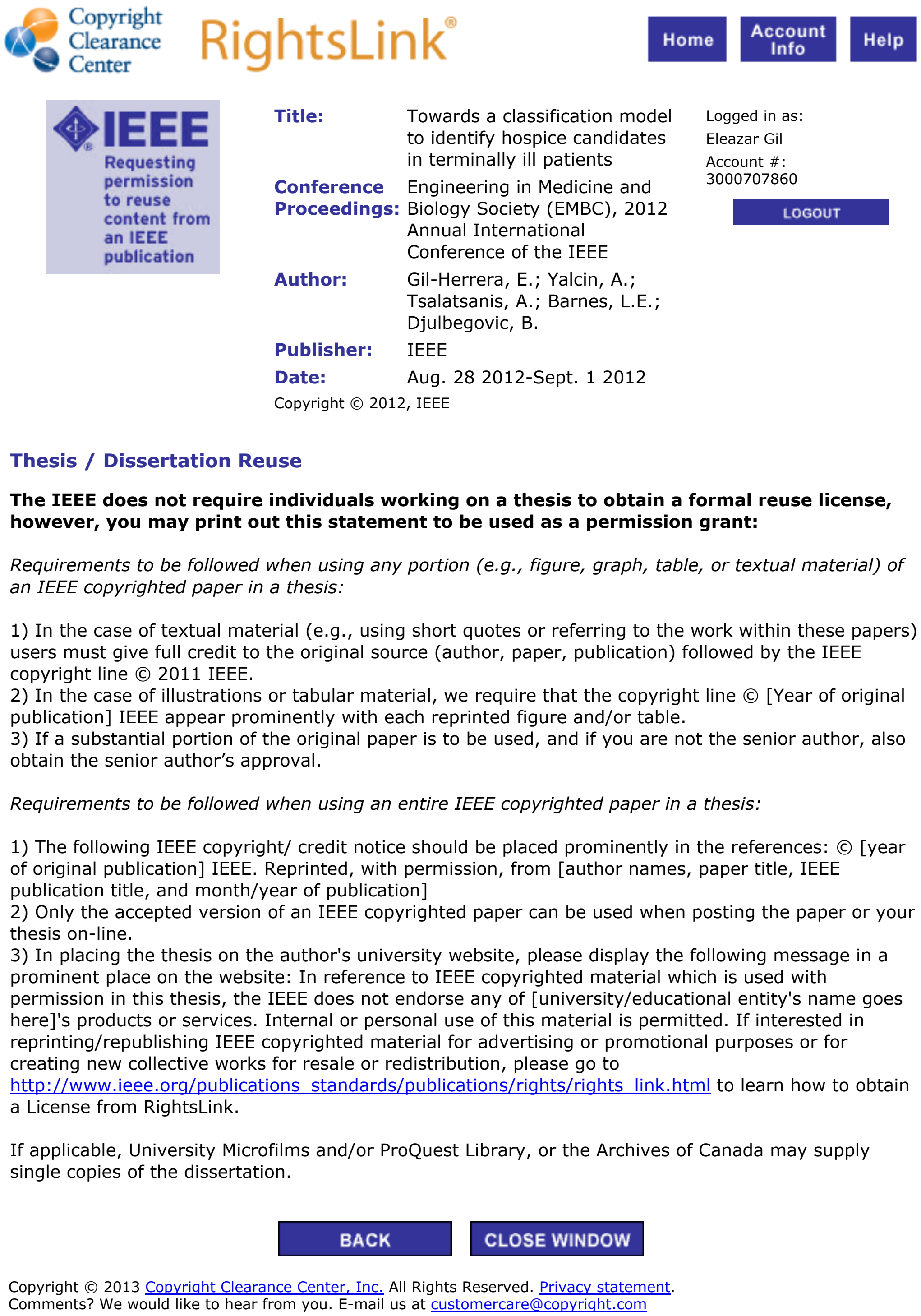
Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Copyright © 2013 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement.

Appendix A (continued)


Copyright Clearance Center
RightsLink®
Home
Account Info
Help
IEEE
Requesting permission to reuse content from an IEEE publication
LOGOUT
BACK
CLOSE WINDOW

Title: Towards a classification model to identify hospice candidates in terminally ill patients

Conference Proceedings: Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE

Author: Gil-Herrera, E.; Yalcin, A.; Tsalatsanis, A.; Barnes, L.E.; Djulbegovic, B.

Publisher: IEEE

Date: Aug. 28 2012-Sept. 1 2012

Copyright © 2012, IEEE

Logged in as:
Eleazar Gil
Account #:
3000707860

Thesis / Dissertation Reuse

The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © 2011 IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author's approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/ credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names, paper title, IEEE publication title, and month/year of publication]
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of [university/educational entity's name goes here]'s products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

Copyright © 2013 Copyright Clearance Center, Inc. All Rights Reserved. Privacy statement.
Comments? We would like to hear from you. E-mail us at customercare@copyright.com

Appendix B: Predicting Academic Performance Using a Rough Set Theory-Based Knowledge Discovery Methodology[1]

[1] Predicting Academic Performance Using a Rough Set Theory-Based Knowledge Discovery Methodology. Published in the International Journal of Engineering Education. vol. 27, No 5, pages 992-1002, 2011

Appendix B (continued)

International Journal of Engineering Education Vol. 27, No. 5, pp. 992–1002, 2011
Printed in Great Britain

0949-149X/91 $3.00+0.00
© 2011 TEMPUS Publications.

Predicting Academic Performance Using a Rough Set Theory-Based Knowledge Discovery Methodology*

ELEAZAR GIL-HERRERA,[1] ATHANASIOS TSALATSANIS,[2] ALI YALCIN[1] and AUTAR KAW[3]

[1] University of South Florida, Industrial and Management Systems Engineering Department, 4202 East Fowler Ave. ENB 118, Tampa, FL 33620 USA. E-mail: eleazar@mail.usf.edu ayalcin@usf.edu
[2] University of South Florida, Center for Evidence-based Medicine and Health Outcomes Research, 12901 Bruce B. Downs Blvd. MDC27, Tampa, FL 33612 USA. E-mail: atsalats@health.usf.edu
[3] University of South Florida, Mechanical Engineering Department, 4202 East Fowler Ave. ENB 118, Tampa, FL 33620 USA. E-mail: kaw@usf.edu

In an effort to predict student performance in an engineering course, Rough Set Theory (RST) is employed as the core of a knowledge discovery process. Student performance is captured in terms of successful course completion. Therefore, students are classified into two categories: those who pass a course and those who do not. The Rough Set Theory paradigm presented here analyzes each student based on a set of attributes. These attributes are collected through a series of surveys conducted in the first week of the course, allowing for early identification of potential unsuccessful students. Variations of the Rough Set approach are evaluated to determine the one most suited for the particular dataset. The results are promising since the accuracy of student performance prediction presents an *Area under the Receiver Operating Characteristic Curve* equal to 80%. The benefits anticipated from early identification of weak and/or potentially unsuccessful students will enable educators to engage these students at the onset of the course and enroll them in additional activities to improve their performance.

Keywords: academic performance prediction; linear systems; rough set theory; knowledge discovery

1. Introduction

Knowledge discovery is the research area concerned with analyzing existing information and extracting implicit, previously unknown, hidden and potentially useful knowledge in an automated manner [1, 2]. The core of the presented knowledge discovery process is Rough Set Theory (RST) [1], an extension to classical Set Theory used to represent incomplete or imperfect knowledge. RST combines theories such as fuzzy sets [3], evidence theory [4] and statistics, hence is able to cope with the shortcomings of these underlying theories.

In this paper, we describe the application of an RST-based knowledge discovery process in predicting student performance in an undergraduate engineering course. We measure student performance in terms of successful completion of a course. In this context, we classify students into two categories: *Passing students* are those who complete the course with a passing grade and *Failing students* are those who fail to complete the course or receive a failing grade. Note that the failing students category is used in the generic sense and includes those students who withdraw from it. The dataset for this study consists of information collected from two distinct groups of students enrolled in two different classes of the course. Student information was collected through a series of surveys conducted in the first week of the classes.

The rest of the paper is organized as follows: Section 2 presents a review of the recent work in predicting student performance in a single course. Section 3 describes the dataset utilized in this study and Section 4 presents in detail each of the steps involved in the RST-based knowledge discovery process and their application to predicting student performance. Section 5 discusses our results and finally, Section 6 concludes this paper.

2. Literature review

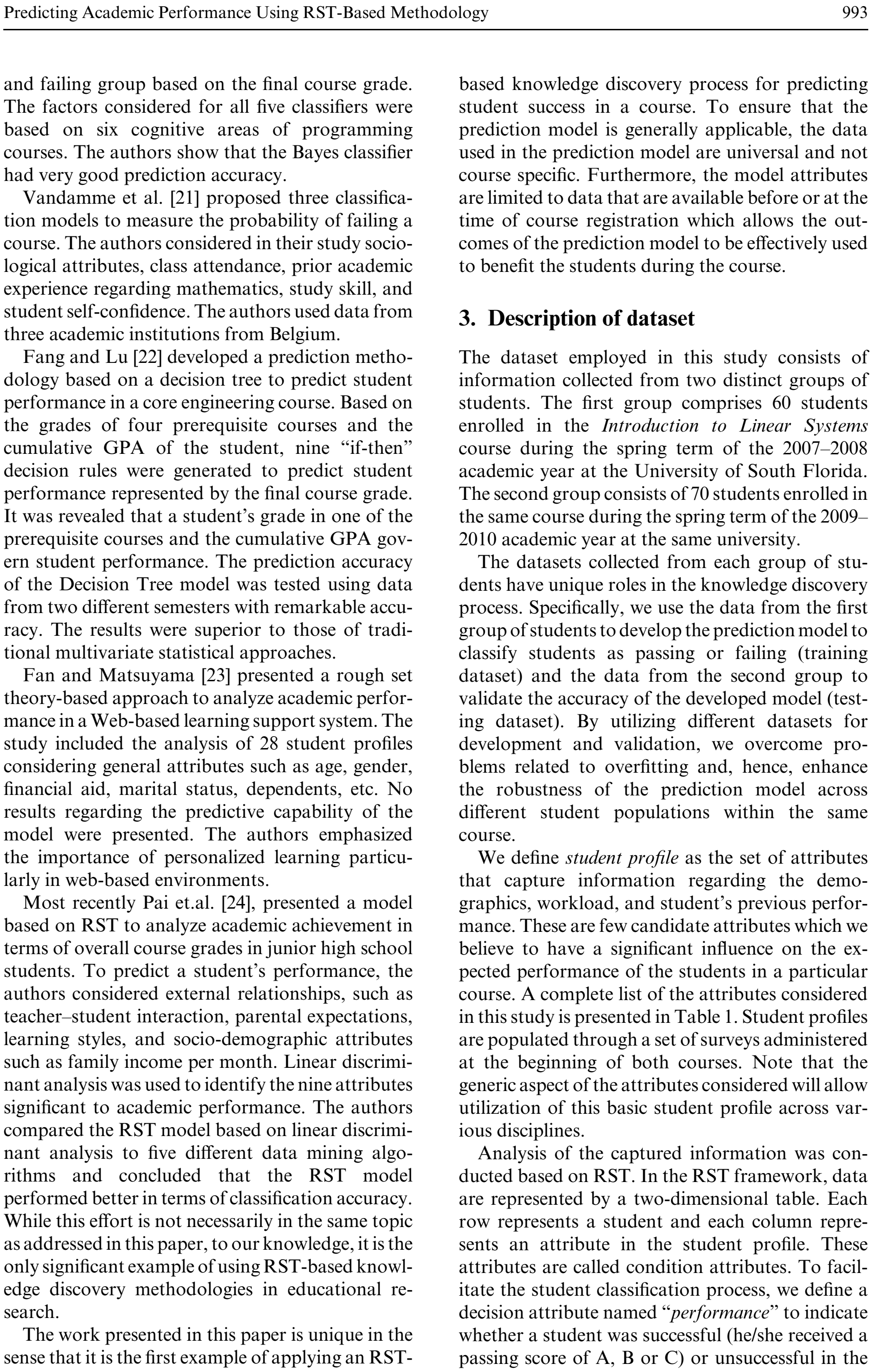
A variety of methodologies has been proposed to predict student performance in academic settings with the majority of them relying on statistical and soft computing techniques. The specific topic of academic performance prediction in a single course is dominated by regression-based statistical approaches. Recent notable efforts based on regression analysis appear in [5–15].

Soft computing techniques have found application in student performance prediction in the broader sense of overall academic success and retention in [16–19]. There are also notable efforts in applying these techniques to student performance prediction in a single course.

Hamalainen and Vinni [20] compared five student performance classification methods; two multiple linear regression and three versions of naïve Bayes classifiers. Students were classified into a passing

* Accepted 16 April 2011.

Appendix B (continued)



and failing group based on the final course grade. The factors considered for all five classifiers were based on six cognitive areas of programming courses. The authors show that the Bayes classifier had very good prediction accuracy.

Vandamme et al. [21] proposed three classification models to measure the probability of failing a course. The authors considered in their study sociological attributes, class attendance, prior academic experience regarding mathematics, study skill, and student self-confidence. The authors used data from three academic institutions from Belgium.

Fang and Lu [22] developed a prediction methodology based on a decision tree to predict student performance in a core engineering course. Based on the grades of four prerequisite courses and the cumulative GPA of the student, nine "if-then" decision rules were generated to predict student performance represented by the final course grade. It was revealed that a student's grade in one of the prerequisite courses and the cumulative GPA govern student performance. The prediction accuracy of the Decision Tree model was tested using data from two different semesters with remarkable accuracy. The results were superior to those of traditional multivariate statistical approaches.

Fan and Matsuyama [23] presented a rough set theory-based approach to analyze academic performance in a Web-based learning support system. The study included the analysis of 28 student profiles considering general attributes such as age, gender, financial aid, marital status, dependents, etc. No results regarding the predictive capability of the model were presented. The authors emphasized the importance of personalized learning particularly in web-based environments.

Most recently Pai et.al. [24], presented a model based on RST to analyze academic achievement in terms of overall course grades in junior high school students. To predict a student's performance, the authors considered external relationships, such as teacher–student interaction, parental expectations, learning styles, and socio-demographic attributes such as family income per month. Linear discriminant analysis was used to identify the nine attributes significant to academic performance. The authors compared the RST model based on linear discriminant analysis to five different data mining algorithms and concluded that the RST model performed better in terms of classification accuracy. While this effort is not necessarily in the same topic as addressed in this paper, to our knowledge, it is the only significant example of using RST-based knowledge discovery methodologies in educational research.

The work presented in this paper is unique in the sense that it is the first example of applying an RST-based knowledge discovery process for predicting student success in a course. To ensure that the prediction model is generally applicable, the data used in the prediction model are universal and not course specific. Furthermore, the model attributes are limited to data that are available before or at the time of course registration which allows the outcomes of the prediction model to be effectively used to benefit the students during the course.

3. Description of dataset

The dataset employed in this study consists of information collected from two distinct groups of students. The first group comprises 60 students enrolled in the *Introduction to Linear Systems* course during the spring term of the 2007–2008 academic year at the University of South Florida. The second group consists of 70 students enrolled in the same course during the spring term of the 2009–2010 academic year at the same university.

The datasets collected from each group of students have unique roles in the knowledge discovery process. Specifically, we use the data from the first group of students to develop the prediction model to classify students as passing or failing (training dataset) and the data from the second group to validate the accuracy of the developed model (testing dataset). By utilizing different datasets for development and validation, we overcome problems related to overfitting and, hence, enhance the robustness of the prediction model across different student populations within the same course.

We define *student profile* as the set of attributes that capture information regarding the demographics, workload, and student's previous performance. These are few candidate attributes which we believe to have a significant influence on the expected performance of the students in a particular course. A complete list of the attributes considered in this study is presented in Table 1. Student profiles are populated through a set of surveys administered at the beginning of both courses. Note that the generic aspect of the attributes considered will allow utilization of this basic student profile across various disciplines.

Analysis of the captured information was conducted based on RST. In the RST framework, data are represented by a two-dimensional table. Each row represents a student and each column represents an attribute in the student profile. These attributes are called condition attributes. To facilitate the student classification process, we define a decision attribute named "*performance*" to indicate whether a student was successful (he/she received a passing score of A, B or C) or unsuccessful in the

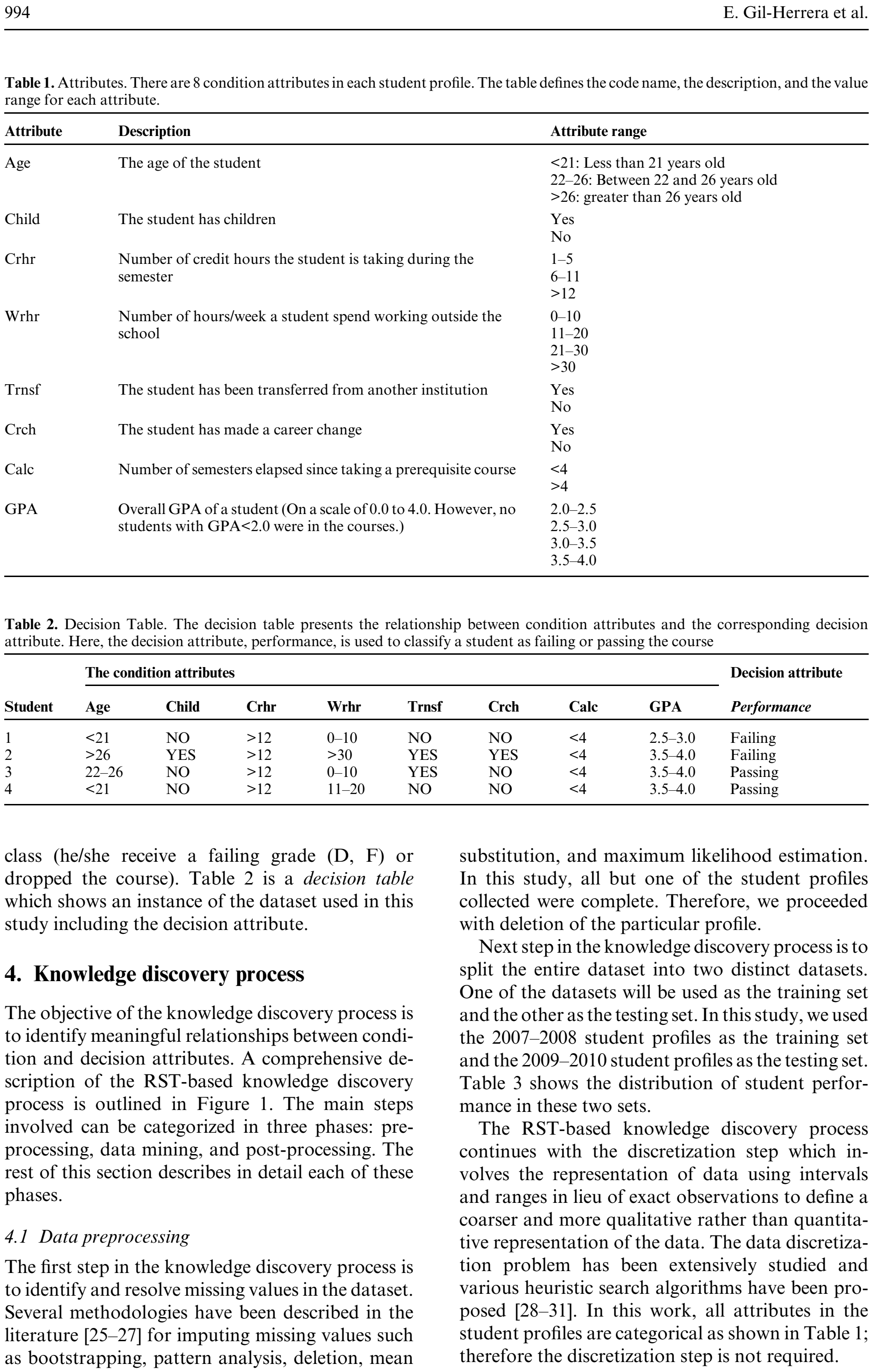


Table 1. Attributes. There are 8 condition attributes in each student profile. The table defines the code name, the description, and the value range for each attribute.

Attribute	Description	Attribute range
Age	The age of the student	<21: Less than 21 years old 22–26: Between 22 and 26 years old >26: greater than 26 years old
Child	The student has children	Yes No
Crhr	Number of credit hours the student is taking during the semester	1–5 6–11 >12
Wrhr	Number of hours/week a student spend working outside the school	0–10 11–20 21–30 >30
Trnsf	The student has been transferred from another institution	Yes No
Crch	The student has made a career change	Yes No
Calc	Number of semesters elapsed since taking a prerequisite course	<4 >4
GPA	Overall GPA of a student (On a scale of 0.0 to 4.0. However, no students with GPA<2.0 were in the courses.)	2.0–2.5 2.5–3.0 3.0–3.5 3.5–4.0

Table 2. Decision Table. The decision table presents the relationship between condition attributes and the corresponding decision attribute. Here, the decision attribute, performance, is used to classify a student as failing or passing the course

	The condition attributes								**Decision attribute**
Student	**Age**	**Child**	**Crhr**	**Wrhr**	**Trnsf**	**Crch**	**Calc**	**GPA**	***Performance***
1	<21	NO	>12	0–10	NO	NO	<4	2.5–3.0	Failing
2	>26	YES	>12	>30	YES	YES	<4	3.5–4.0	Failing
3	22–26	NO	>12	0–10	YES	NO	<4	3.5–4.0	Passing
4	<21	NO	>12	11–20	NO	NO	<4	3.5–4.0	Passing

class (he/she receive a failing grade (D, F) or dropped the course). Table 2 is a *decision table* which shows an instance of the dataset used in this study including the decision attribute.

4. Knowledge discovery process

The objective of the knowledge discovery process is to identify meaningful relationships between condition and decision attributes. A comprehensive description of the RST-based knowledge discovery process is outlined in Figure 1. The main steps involved can be categorized in three phases: preprocessing, data mining, and post-processing. The rest of this section describes in detail each of these phases.

4.1 Data preprocessing

The first step in the knowledge discovery process is to identify and resolve missing values in the dataset. Several methodologies have been described in the literature [25–27] for imputing missing values such as bootstrapping, pattern analysis, deletion, mean substitution, and maximum likelihood estimation. In this study, all but one of the student profiles collected were complete. Therefore, we proceeded with deletion of the particular profile.

Next step in the knowledge discovery process is to split the entire dataset into two distinct datasets. One of the datasets will be used as the training set and the other as the testing set. In this study, we used the 2007–2008 student profiles as the training set and the 2009–2010 student profiles as the testing set. Table 3 shows the distribution of student performance in these two sets.

The RST-based knowledge discovery process continues with the discretization step which involves the representation of data using intervals and ranges in lieu of exact observations to define a coarser and more qualitative rather than quantitative representation of the data. The data discretization problem has been extensively studied and various heuristic search algorithms have been proposed [28–31]. In this work, all attributes in the student profiles are categorical as shown in Table 1; therefore the discretization step is not required.

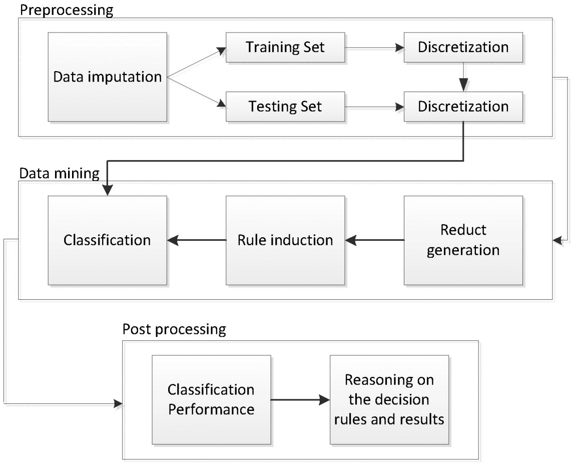


Fig. 1. Knowledge discovery methodology. There are three phases in the knowledge discovery process: data preprocessing, data mining, and data post processing.

Table 3. Performance distribution in training and testing sets

Dataset	Failing	Passing
Training set	58.33%	41.67%
Testing set	37.68%	62.32%

4.2 Reduct generation

The reduct generation step is utilized in an effort to reduce the dimensionality of the dataset by removing redundant information and consequently decreasing the complexity of the mining process. Formally, a reduct is the minimal set of attributes that enable the same classification as the complete set of attributes without loss of information. There are many algorithms for computing reducts. As will be shown later in this paper, the effect of the reduct generation algorithm to the classification performance is critical. Therefore, the optimal algorithm is identified as the one producing the best classification results. However, since the computational complexity of the reduct generation problem is NP-hard [28, 32], various suboptimal techniques have been proposed. The technique most appropriate to the problem is the one that generates better classification accuracy in the testing dataset. In this work, two techniques are used for reduct generation: genetic algorithms and dynamic reducts. The rest of this section describes these techniques.

4.2.1 Computing reducts using genetic algorithms

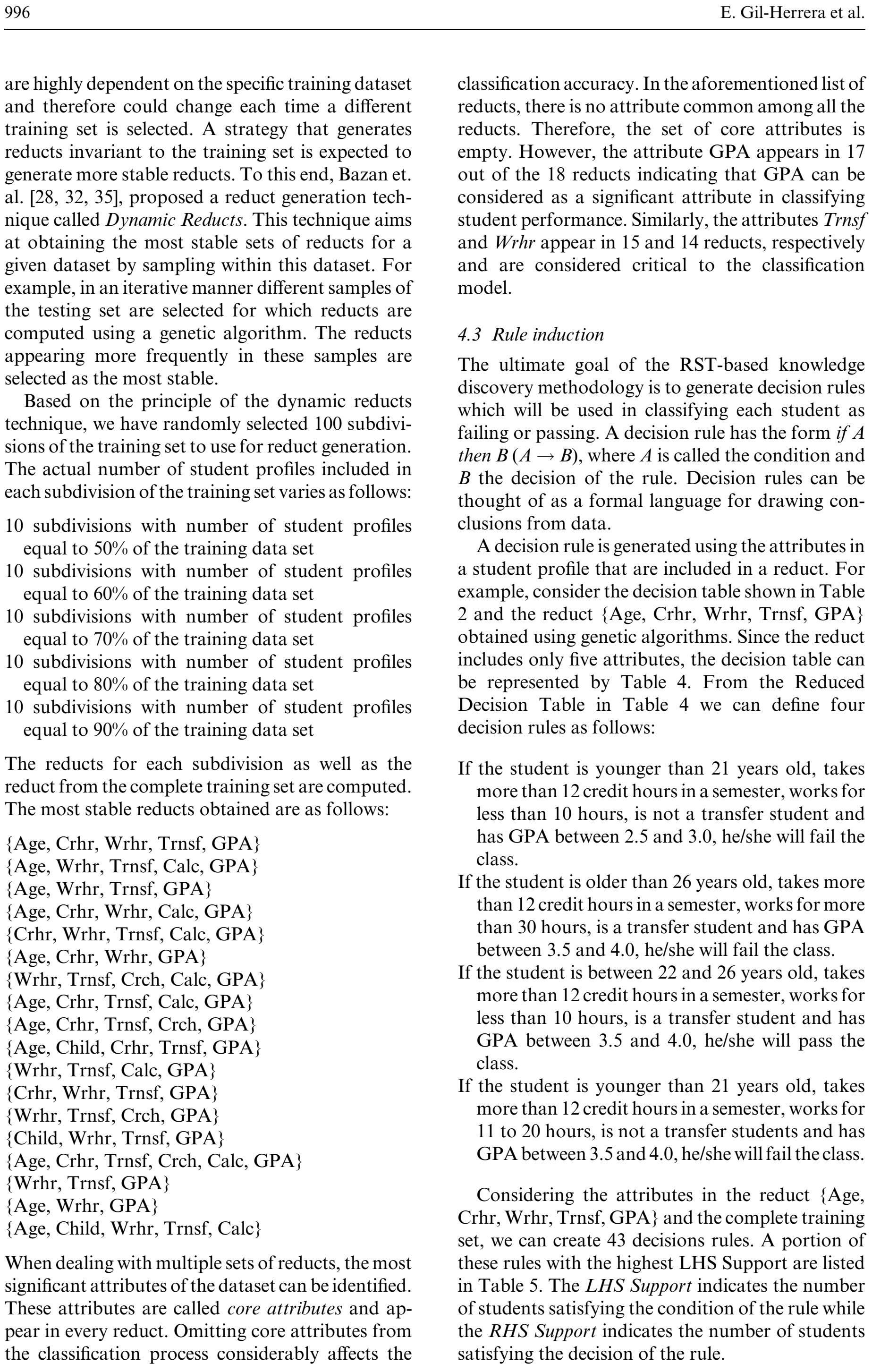
The computational cost for reduct computation is exponential with respect to the size of the decision table. Genetic algorithms, operating based on the principle of survival of the fittest, can be used to reduce the computational complexity [32–34]. Given a function $f : S_{+}$, the goal of a genetic algorithm is to find an $x_0 \in S$ for which $f(x_0) = \max(f(x) : x \in S)$. Elements of S are called *individuals* and the function f is the *fitness function*. The values of function $f(x)$ correspond to the ability of the individual x to survive the evolution process. The evolution process begins by creating a random initial fixed size population of individuals. In an iterative manner, the algorithm generates a new population of individuals. First, the fitness of each individual in the current population is calculated and those individuals with high fitness are selected as parents which interact based on a genetic operator (e.g. mutation and crossover) to produce the new population, child. The process is repeated until some stopping condition is achieved.

The genetic algorithm for the reduct generation uses as *individuals* the attributes in the student profile, and as fitness function the output of a heuristic algorithm that evaluates the quality of each reduct generated. The details of the genetic algorithm used for the reduct generation are presented in [32]. Using genetic algorithms, one reduct {Age, Crhr, Wrhr, Trnsf, GPA} is generated which includes 5 out of the 8 attributes.

4.2.2 Computing dynamic reducts

The main advantage of utilizing genetic algorithms for reduct generation is the reduction in computational complexity. However, the results obtained

Appendix B (continued)



are highly dependent on the specific training dataset and therefore could change each time a different training set is selected. A strategy that generates reducts invariant to the training set is expected to generate more stable reducts. To this end, Bazan et. al. [28, 32, 35], proposed a reduct generation technique called *Dynamic Reducts*. This technique aims at obtaining the most stable sets of reducts for a given dataset by sampling within this dataset. For example, in an iterative manner different samples of the testing set are selected for which reducts are computed using a genetic algorithm. The reducts appearing more frequently in these samples are selected as the most stable.

Based on the principle of the dynamic reducts technique, we have randomly selected 100 subdivisions of the training set to use for reduct generation. The actual number of student profiles included in each subdivision of the training set varies as follows:

10 subdivisions with number of student profiles equal to 50% of the training data set
10 subdivisions with number of student profiles equal to 60% of the training data set
10 subdivisions with number of student profiles equal to 70% of the training data set
10 subdivisions with number of student profiles equal to 80% of the training data set
10 subdivisions with number of student profiles equal to 90% of the training data set

The reducts for each subdivision as well as the reduct from the complete training set are computed. The most stable reducts obtained are as follows:

{Age, Crhr, Wrhr, Trnsf, GPA}
{Age, Wrhr, Trnsf, Calc, GPA}
{Age, Wrhr, Trnsf, GPA}
{Age, Crhr, Wrhr, Calc, GPA}
{Crhr, Wrhr, Trnsf, Calc, GPA}
{Age, Crhr, Wrhr, GPA}
{Wrhr, Trnsf, Crch, Calc, GPA}
{Age, Crhr, Trnsf, Calc, GPA}
{Age, Crhr, Trnsf, Crch, GPA}
{Age, Child, Crhr, Trnsf, GPA}
{Wrhr, Trnsf, Calc, GPA}
{Crhr, Wrhr, Trnsf, GPA}
{Wrhr, Trnsf, Crch, GPA}
{Child, Wrhr, Trnsf, GPA}
{Age, Crhr, Trnsf, Crch, Calc, GPA}
{Wrhr, Trnsf, GPA}
{Age, Wrhr, GPA}
{Age, Child, Wrhr, Trnsf, Calc}

When dealing with multiple sets of reducts, the most significant attributes of the dataset can be identified. These attributes are called *core attributes* and appear in every reduct. Omitting core attributes from the classification process considerably affects the classification accuracy. In the aforementioned list of reducts, there is no attribute common among all the reducts. Therefore, the set of core attributes is empty. However, the attribute GPA appears in 17 out of the 18 reducts indicating that GPA can be considered as a significant attribute in classifying student performance. Similarly, the attributes *Trnsf* and *Wrhr* appear in 15 and 14 reducts, respectively and are considered critical to the classification model.

4.3 Rule induction

The ultimate goal of the RST-based knowledge discovery methodology is to generate decision rules which will be used in classifying each student as failing or passing. A decision rule has the form *if A then B* ($A \rightarrow B$), where A is called the condition and B the decision of the rule. Decision rules can be thought of as a formal language for drawing conclusions from data.

A decision rule is generated using the attributes in a student profile that are included in a reduct. For example, consider the decision table shown in Table 2 and the reduct {Age, Crhr, Wrhr, Trnsf, GPA} obtained using genetic algorithms. Since the reduct includes only five attributes, the decision table can be represented by Table 4. From the Reduced Decision Table in Table 4 we can define four decision rules as follows:

If the student is younger than 21 years old, takes more than 12 credit hours in a semester, works for less than 10 hours, is not a transfer student and has GPA between 2.5 and 3.0, he/she will fail the class.
If the student is older than 26 years old, takes more than 12 credit hours in a semester, works for more than 30 hours, is a transfer student and has GPA between 3.5 and 4.0, he/she will fail the class.
If the student is between 22 and 26 years old, takes more than 12 credit hours in a semester, works for less than 10 hours, is a transfer student and has GPA between 3.5 and 4.0, he/she will pass the class.
If the student is younger than 21 years old, takes more than 12 credit hours in a semester, works for 11 to 20 hours, is not a transfer students and has GPA between 3.5 and 4.0, he/she will fail the class.

Considering the attributes in the reduct {Age, Crhr, Wrhr, Trnsf, GPA} and the complete training set, we can create 43 decisions rules. A portion of these rules with the highest LHS Support are listed in Table 5. The *LHS Support* indicates the number of students satisfying the condition of the rule while the *RHS Support* indicates the number of students satisfying the decision of the rule.

Appendix B (continued)

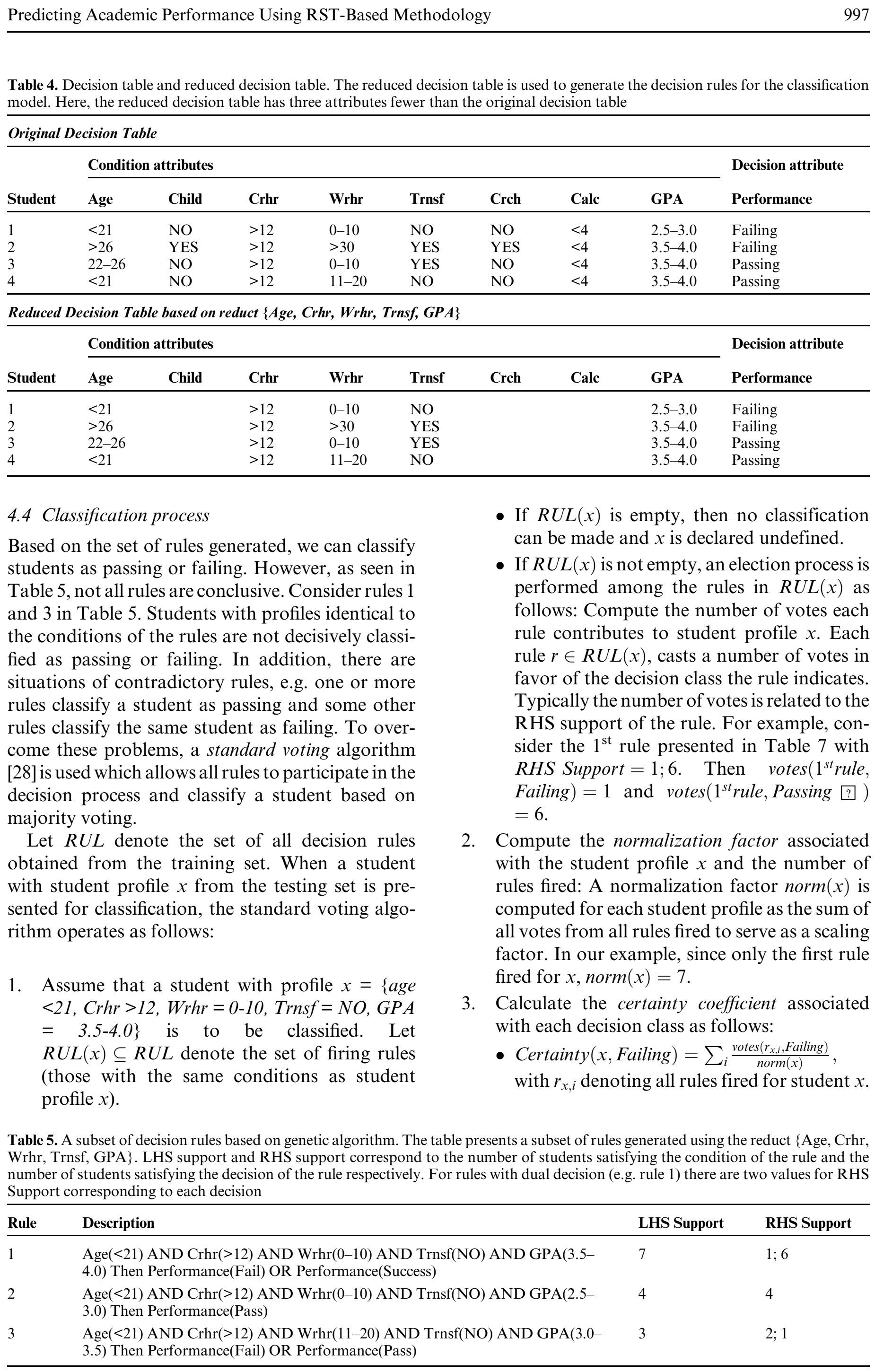


Table 4. Decision table and reduced decision table. The reduced decision table is used to generate the decision rules for the classification model. Here, the reduced decision table has three attributes fewer than the original decision table

Original Decision Table

	Condition attributes								Decision attribute
Student	Age	Child	Crhr	Wrhr	Trnsf	Crch	Calc	GPA	Performance
1	<21	NO	>12	0–10	NO	NO	<4	2.5–3.0	Failing
2	>26	YES	>12	>30	YES	YES	<4	3.5–4.0	Failing
3	22–26	NO	>12	0–10	YES	NO	<4	3.5–4.0	Passing
4	<21	NO	>12	11–20	NO	NO	<4	3.5–4.0	Passing

Reduced Decision Table based on reduct {Age, Crhr, Wrhr, Trnsf, GPA}

	Condition attributes								Decision attribute
Student	Age	Child	Crhr	Wrhr	Trnsf	Crch	Calc	GPA	Performance
1	<21		>12	0–10	NO			2.5–3.0	Failing
2	>26		>12	>30	YES			3.5–4.0	Failing
3	22–26		>12	0–10	YES			3.5–4.0	Passing
4	<21		>12	11–20	NO			3.5–4.0	Passing

4.4 Classification process

Based on the set of rules generated, we can classify students as passing or failing. However, as seen in Table 5, not all rules are conclusive. Consider rules 1 and 3 in Table 5. Students with profiles identical to the conditions of the rules are not decisively classified as passing or failing. In addition, there are situations of contradictory rules, e.g. one or more rules classify a student as passing and some other rules classify the same student as failing. To overcome these problems, a *standard voting* algorithm [28] is used which allows all rules to participate in the decision process and classify a student based on majority voting.

Let RUL denote the set of all decision rules obtained from the training set. When a student with student profile x from the testing set is presented for classification, the standard voting algorithm operates as follows:

1. Assume that a student with profile x = {*age <21, Crhr >12, Wrhr = 0-10, Trnsf = NO, GPA = 3.5-4.0*} is to be classified. Let $RUL(x) \subseteq RUL$ denote the set of firing rules (those with the same conditions as student profile x).
   - If $RUL(x)$ is empty, then no classification can be made and x is declared undefined.
   - If $RUL(x)$ is not empty, an election process is performed among the rules in $RUL(x)$ as follows: Compute the number of votes each rule contributes to student profile x. Each rule $r \in RUL(x)$, casts a number of votes in favor of the decision class the rule indicates. Typically the number of votes is related to the RHS support of the rule. For example, consider the 1st rule presented in Table 7 with $RHS\ Support = 1; 6$. Then $votes(1^{st} rule, Failing) = 1$ and $votes(1^{st} rule, Passing\ \boxed{?}\ ) = 6$.
2. Compute the *normalization factor* associated with the student profile x and the number of rules fired: A normalization factor $norm(x)$ is computed for each student profile as the sum of all votes from all rules fired to serve as a scaling factor. In our example, since only the first rule fired for x, $norm(x) = 7$.
3. Calculate the *certainty coefficient* associated with each decision class as follows:
   - $Certainty(x, Failing) = \sum_i \frac{votes(r_{x,i}, Failing)}{norm(x)}$, with $r_{x,i}$ denoting all rules fired for student x.

Table 5. A subset of decision rules based on genetic algorithm. The table presents a subset of rules generated using the reduct {Age, Crhr, Wrhr, Trnsf, GPA}. LHS support and RHS support correspond to the number of students satisfying the condition of the rule and the number of students satisfying the decision of the rule respectively. For rules with dual decision (e.g. rule 1) there are two values for RHS Support corresponding to each decision

Rule	Description	LHS Support	RHS Support
1	Age(<21) AND Crhr(>12) AND Wrhr(0–10) AND Trnsf(NO) AND GPA(3.5–4.0) Then Performance(Fail) OR Performance(Success)	7	1; 6
2	Age(<21) AND Crhr(>12) AND Wrhr(0–10) AND Trnsf(NO) AND GPA(2.5–3.0) Then Performance(Pass)	4	4
3	Age(<21) AND Crhr(>12) AND Wrhr(11–20) AND Trnsf(NO) AND GPA(3.0–3.5) Then Performance(Fail) OR Performance(Pass)	3	2; 1

Appendix B (continued)

- $Certainty(x, Passing) = \sum_i \frac{votes(r_{x,i}, Passing)}{norm(x)}$. For our example, $Certainty(x, Failing) = \frac{1}{7}$ and $Certainty(x, Passing) = \frac{6}{7}$.

4. Finally, classify the student with profile x in the decision class for which the certainty factor is greater than a *threshold value* (τ) which is typically fixed at 0.5. In this example, the student with profile x is classified as *Passing*.

5. Results

This section compares the performance of the classification processes based on the decision rules generated using the reduct generation techniques described in sections 4.2.1–4.2.2. At this stage of the knowledge discovery methodology, the objects (student profiles) in training dataset are classified as passing, failing or undefined based on the induced rules and the classification process described. The results are presented in a confusion matrix form. The confusion matrix for each model includes the numbers of True Positive (*TP*), True Negative (*TN*), False Positive (*FP*) and False Negative (*FN*) results. Our perspective on positive and negative results relates to the necessitation for action for failing students. Specifically, we define:

TP: the number of students classified as failing the course, when in fact failed the course (shown in the top left cell of the confusion matrix).

FP: the number of students classified as failing the course, when in fact passed the course (shown in the bottom left cell of the confusion matrix).

TN: the number of students classified as passing the course, when in fact passed the course (shown in the bottom right cell of the confusion matrix).

FN: the number of students classified as passing the course, when in fact failed the course (shown in the top right cell of the confusion matrix).

Using these values we can compute the measures of specificity and sensitivity as:

Sensitivity: The fraction of failing students correctly classified by the classification algorithm.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

Specificity: The fraction of passing students correctly classified by the classification algorithm.

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

The accuracy of each classification model is reported in terms of Area under the Receiver Operating Characteristic (ROC) curve (AUC). The ROC curve graphs the sensitivity of the classification algorithm in terms of (1-specificity). The best possible classification is achieved when AUC is equal to 1, while no classification ability exists when AUC is equal to 0.5.

5.1 Performance of the classification algorithm using reducts generated by genetic algorithms

Table 6 presents the confusion matrix for the classification model based on reducts generated using genetic algorithms. The classifier consists of 43 rules. With sensitivity equal to 80%, the classifier demonstrates an ability to correctly identify the failing students, however, the specificity score is much lower (20%), which implies that the classifier fails to correctly identify passing students. The term *undefined* in Table 6 refers to 59 students (almost 85.5% of students in the testing sample) for whom the classification algorithm was unable to classify either as passing or failing. The coverage of the classifier (defined by the ratio of objects classified to the total number of objects in the testing set) is 14.5% since we are able to classify 10 students from the 70 in the training set. Overall, the AUC score is equal to 0.5 indicating classification inability.

5.2 Performance of the classification algorithm using dynamic reducts

Table 7 shows the confusion matrix for the classification model based on dynamic reducts. There are 593 decision rules. The classifier's ability to correctly identify failing and passing students is 0.68 and 0.675, respectively. The overall classification performance as indicated by the AUC is equal to 0.8, considerably better compared to the genetic algorithm classifier. In addition, the number of undefined cases has been decreased to four student profiles and the coverage of the classifier is 96%. Using dynamic reducts instead of genetic algo-

Table 6. Confusion matrix. The classifier presents AUC equal to 0.5 indicating classification inability

	Predicted		
	Failing	**Passing**	**Undefined**
Actual			
Failing	4	1	21
Passing	4	1	38
Sensitivity: 0.8, Specificity: 0.2, AUC: 0.5			

Table 7. Confusion matrix. The classifier presents AUC equal to 0.8 indicates good classification ability

	Predicted		
	Failing	**Passing**	**Undefined**
Actual			
Failing	17	8	1
Passing	13	27	3
Sensitivity: 0.68, Specificity: 0.675, AUC: 0.8			

Appendix B (continued)

Table 8. Comparison of classifiers. A classifier has been created based on each reduct generation technique described in sections 4.2.1–4.2.2

	Strategy	
Performance measures	**Genetic Algorithms**	**Dynamic reducts**
Sensitivity	0.8	0.68
Specificity	0.2	0.675
AUC	0.5	0.8
Coverage	14.5 %	94%
# of reducts	1	18
# of decision rules	43	593

rithms for reduct generation improved the overall classification performance.

Table 8 summarizes our findings regarding the performance of each classifier in predicting student performance.

6. Discussion

The *threshold value* (τ) in the classification process described in Section 4.4 has a significant impact on the accuracy as well as the usability of the classification process, especially in this application of student performance prediction. To better understand the role of this threshold value, consider the definitions of sensitivity, the fraction of failing students correctly classified, and specificity, the fraction of passing students correctly classified by the classification algorithm. In our particular application of predicting student performance in a course, to engage the potentially unsuccessful students early on and to improve their performance, the "cost" of misclassifying a failing student (as passing) is much higher than that of misclassifying a passing student (as failing). After all, if a potentially weak/unsuccessful student is misclassified as passing, the opportunity to engage this student early is lost. On the other hand, if a passing student is misclassified as failing and is enrolled in activities to improve his/her performance, he/she may actually end up with an improved grade. Therefore, especially in this particular application, it is significantly more important to ensure that the sensitivity value is closer to 1 than the specificity value.

The threshold value is the parameter that establishes the relation between sensitivity and specificity in the classification process. A higher threshold value would require a higher certainty coefficient value (making it more difficult) for a student to be classified as failing, decreasing the sensitivity and increasing specificity. In the same manner, a lower threshold value would increase sensitivity and reduce specificity, which is the more desirable condition in this application.

The ROC curve describes the predictive behavior of a classifier for varying values of the threshold ($0 \leq \tau \leq 1$), in terms of sensitivity, specificity and classifier accuracy. Figure 2 shows the ROC curve generated from the classification model based on dynamic reducts. The area under the ROC curve characterizes the overall accuracy of the classifier. Each point on the curve corresponds to a different pair of sensitivity and specificity values based on varying the value of the threshold (τ).

Table 9 shows some selected points on the ROC curve and the associated threshold value used during the classification process. For example, the default value of $\tau = 0.5$ leads to the sensitivity and specificity values reported in Table 7. The conditional maximum values of both sensitivity and specificity are obtained when the threshold values

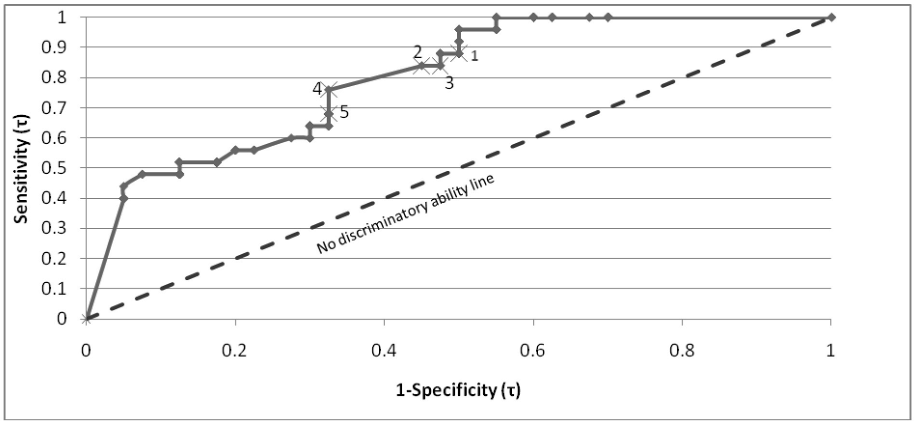


Fig. 2. ROC Curve for the classification model based on dynamic reducts. The specificity and sensitivity are controlled by the threshold value.

Appendix B (continued)

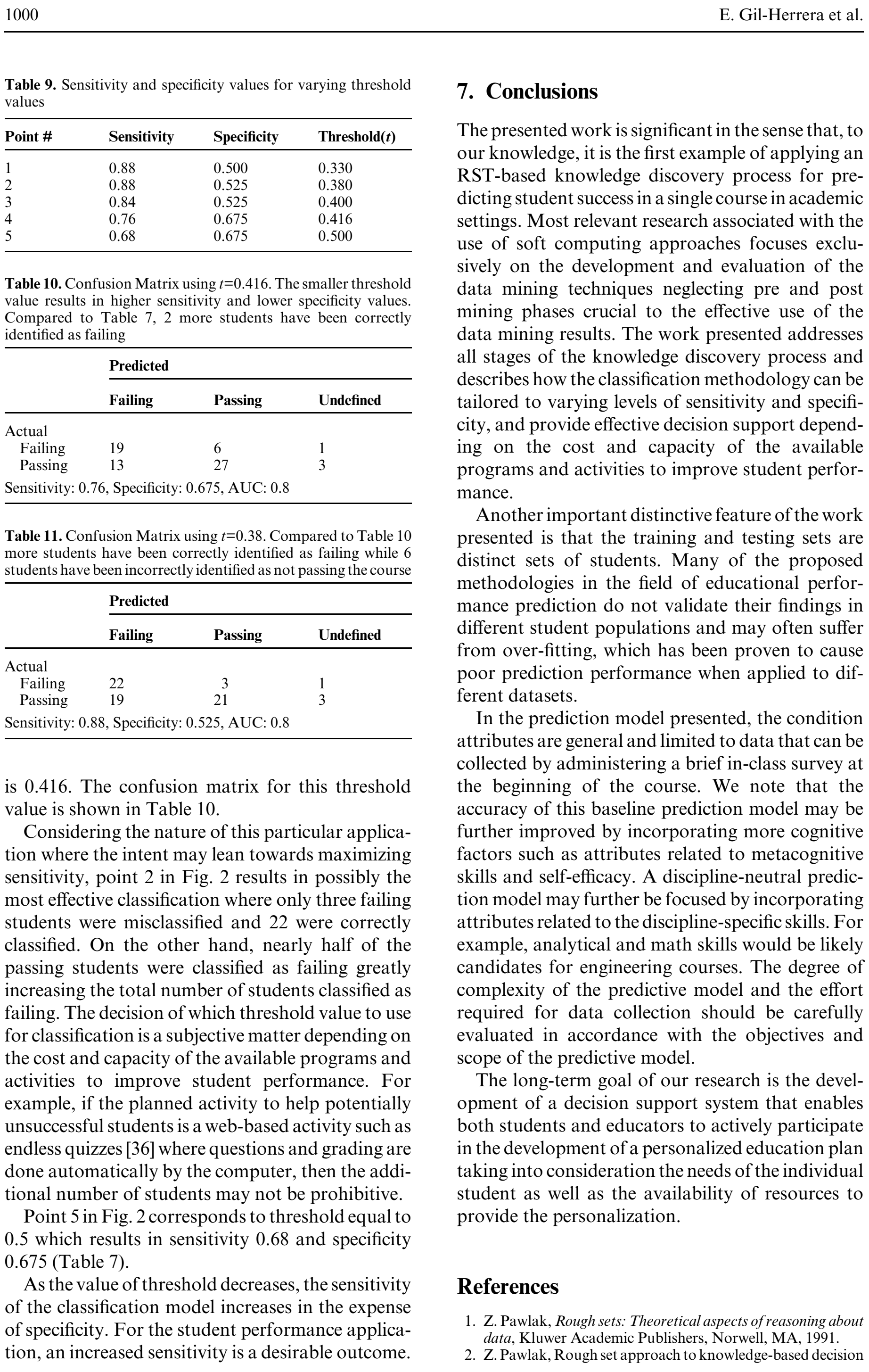


Table 9. Sensitivity and specificity values for varying threshold values

Point #	Sensitivity	Specificity	Threshold(t)
1	0.88	0.500	0.330
2	0.88	0.525	0.380
3	0.84	0.525	0.400
4	0.76	0.675	0.416
5	0.68	0.675	0.500

Table 10. Confusion Matrix using t=0.416. The smaller threshold value results in higher sensitivity and lower specificity values. Compared to Table 7, 2 more students have been correctly identified as failing

	Predicted		
	Failing	Passing	Undefined
Actual			
Failing	19	6	1
Passing	13	27	3
Sensitivity: 0.76, Specificity: 0.675, AUC: 0.8			

Table 11. Confusion Matrix using t=0.38. Compared to Table 10 more students have been correctly identified as failing while 6 students have been incorrectly identified as not passing the course

	Predicted		
	Failing	Passing	Undefined
Actual			
Failing	22	3	1
Passing	19	21	3
Sensitivity: 0.88, Specificity: 0.525, AUC: 0.8			

is 0.416. The confusion matrix for this threshold value is shown in Table 10.

Considering the nature of this particular application where the intent may lean towards maximizing sensitivity, point 2 in Fig. 2 results in possibly the most effective classification where only three failing students were misclassified and 22 were correctly classified. On the other hand, nearly half of the passing students were classified as failing greatly increasing the total number of students classified as failing. The decision of which threshold value to use for classification is a subjective matter depending on the cost and capacity of the available programs and activities to improve student performance. For example, if the planned activity to help potentially unsuccessful students is a web-based activity such as endless quizzes [36] where questions and grading are done automatically by the computer, then the additional number of students may not be prohibitive.

Point 5 in Fig. 2 corresponds to threshold equal to 0.5 which results in sensitivity 0.68 and specificity 0.675 (Table 7).

As the value of threshold decreases, the sensitivity of the classification model increases in the expense of specificity. For the student performance application, an increased sensitivity is a desirable outcome.

7. Conclusions

The presented work is significant in the sense that, to our knowledge, it is the first example of applying an RST-based knowledge discovery process for predicting student success in a single course in academic settings. Most relevant research associated with the use of soft computing approaches focuses exclusively on the development and evaluation of the data mining techniques neglecting pre and post mining phases crucial to the effective use of the data mining results. The work presented addresses all stages of the knowledge discovery process and describes how the classification methodology can be tailored to varying levels of sensitivity and specificity, and provide effective decision support depending on the cost and capacity of the available programs and activities to improve student performance.

Another important distinctive feature of the work presented is that the training and testing sets are distinct sets of students. Many of the proposed methodologies in the field of educational performance prediction do not validate their findings in different student populations and may often suffer from over-fitting, which has been proven to cause poor prediction performance when applied to different datasets.

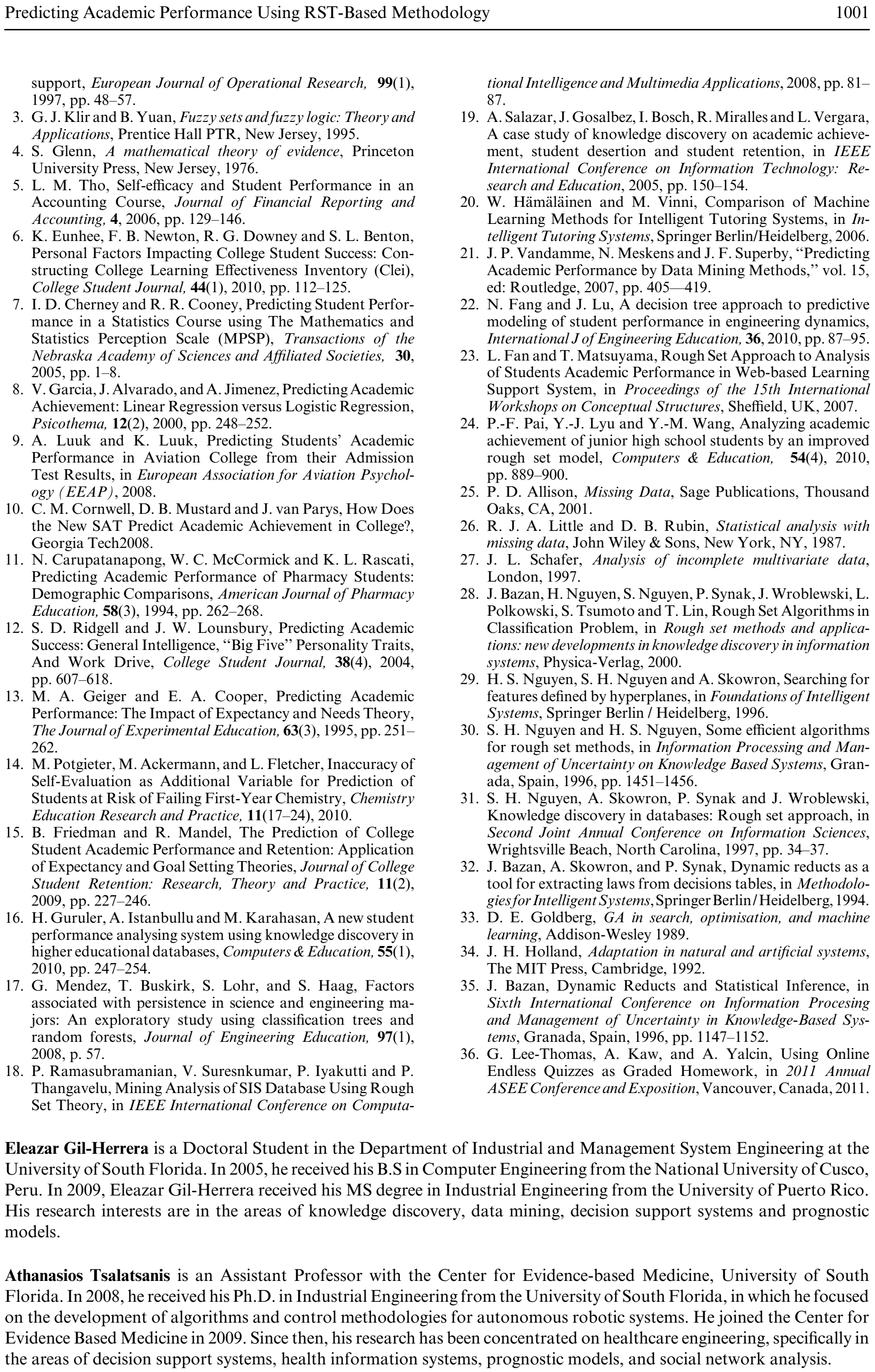
In the prediction model presented, the condition attributes are general and limited to data that can be collected by administering a brief in-class survey at the beginning of the course. We note that the accuracy of this baseline prediction model may be further improved by incorporating more cognitive factors such as attributes related to metacognitive skills and self-efficacy. A discipline-neutral prediction model may further be focused by incorporating attributes related to the discipline-specific skills. For example, analytical and math skills would be likely candidates for engineering courses. The degree of complexity of the predictive model and the effort required for data collection should be carefully evaluated in accordance with the objectives and scope of the predictive model.

The long-term goal of our research is the development of a decision support system that enables both students and educators to actively participate in the development of a personalized education plan taking into consideration the needs of the individual student as well as the availability of resources to provide the personalization.

References

1. Z. Pawlak, *Rough sets: Theoretical aspects of reasoning about data*, Kluwer Academic Publishers, Norwell, MA, 1991.
2. Z. Pawlak, Rough set approach to knowledge-based decision

Appendix B (continued)



Predicting Academic Performance Using RST-Based Methodology 1001

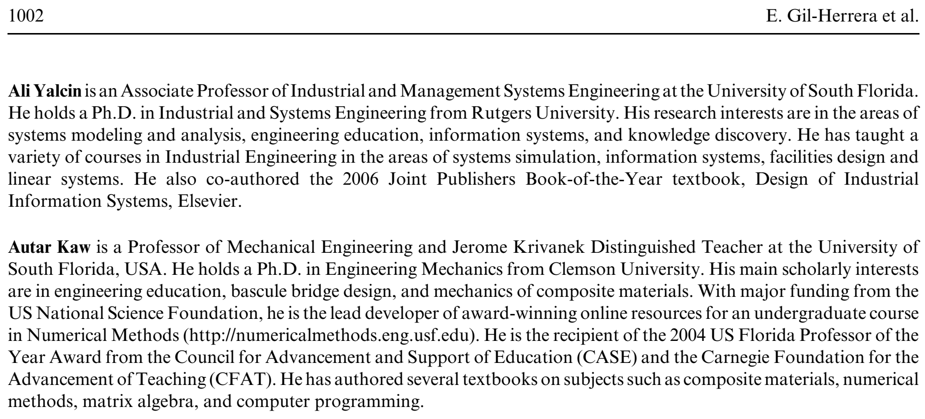
support, *European Journal of Operational Research,* **99**(1), 1997, pp. 48–57.

3. G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: Theory and Applications*, Prentice Hall PTR, New Jersey, 1995.
4. S. Glenn, *A mathematical theory of evidence*, Princeton University Press, New Jersey, 1976.
5. L. M. Tho, Self-efficacy and Student Performance in an Accounting Course, *Journal of Financial Reporting and Accounting,* **4**, 2006, pp. 129–146.
6. K. Eunhee, F. B. Newton, R. G. Downey and S. L. Benton, Personal Factors Impacting College Student Success: Constructing College Learning Effectiveness Inventory (Clei), *College Student Journal,* **44**(1), 2010, pp. 112–125.
7. I. D. Cherney and R. R. Cooney, Predicting Student Performance in a Statistics Course using The Mathematics and Statistics Perception Scale (MPSP), *Transactions of the Nebraska Academy of Sciences and Affiliated Societies,* **30**, 2005, pp. 1–8.
8. V. Garcia, J. Alvarado, and A. Jimenez, Predicting Academic Achievement: Linear Regression versus Logistic Regression, *Psicothema,* **12**(2), 2000, pp. 248–252.
9. A. Luuk and K. Luuk, Predicting Students' Academic Performance in Aviation College from their Admission Test Results, in *European Association for Aviation Psychology (EEAP)*, 2008.
10. C. M. Cornwell, D. B. Mustard and J. van Parys, How Does the New SAT Predict Academic Achievement in College?, Georgia Tech2008.
11. N. Carupatanapong, W. C. McCormick and K. L. Rascati, Predicting Academic Performance of Pharmacy Students: Demographic Comparisons, *American Journal of Pharmacy Education,* **58**(3), 1994, pp. 262–268.
12. S. D. Ridgell and J. W. Lounsbury, Predicting Academic Success: General Intelligence, "Big Five" Personality Traits, And Work Drive, *College Student Journal,* **38**(4), 2004, pp. 607–618.
13. M. A. Geiger and E. A. Cooper, Predicting Academic Performance: The Impact of Expectancy and Needs Theory, *The Journal of Experimental Education,* **63**(3), 1995, pp. 251–262.
14. M. Potgieter, M. Ackermann, and L. Fletcher, Inaccuracy of Self-Evaluation as Additional Variable for Prediction of Students at Risk of Failing First-Year Chemistry, *Chemistry Education Research and Practice,* **11**(17–24), 2010.
15. B. Friedman and R. Mandel, The Prediction of College Student Academic Performance and Retention: Application of Expectancy and Goal Setting Theories, *Journal of College Student Retention: Research, Theory and Practice,* **11**(2), 2009, pp. 227–246.
16. H. Guruler, A. Istanbullu and M. Karahasan, A new student performance analysing system using knowledge discovery in higher educational databases, *Computers & Education,* **55**(1), 2010, pp. 247–254.
17. G. Mendez, T. Buskirk, S. Lohr, and S. Haag, Factors associated with persistence in science and engineering majors: An exploratory study using classification trees and random forests, *Journal of Engineering Education,* **97**(1), 2008, p. 57.
18. P. Ramasubramanian, V. Suresnkumar, P. Iyakutti and P. Thangavelu, Mining Analysis of SIS Database Using Rough Set Theory, in *IEEE International Conference on Computational Intelligence and Multimedia Applications*, 2008, pp. 81–87.
19. A. Salazar, J. Gosalbez, I. Bosch, R. Miralles and L. Vergara, A case study of knowledge discovery on academic achievement, student desertion and student retention, in *IEEE International Conference on Information Technology: Research and Education*, 2005, pp. 150–154.
20. W. Hämäläinen and M. Vinni, Comparison of Machine Learning Methods for Intelligent Tutoring Systems, in *Intelligent Tutoring Systems*, Springer Berlin/Heidelberg, 2006.
21. J. P. Vandamme, N. Meskens and J. F. Superby, "Predicting Academic Performance by Data Mining Methods," vol. 15, ed: Routledge, 2007, pp. 405—419.
22. N. Fang and J. Lu, A decision tree approach to predictive modeling of student performance in engineering dynamics, *International J of Engineering Education,* **36**, 2010, pp. 87–95.
23. L. Fan and T. Matsuyama, Rough Set Approach to Analysis of Students Academic Performance in Web-based Learning Support System, in *Proceedings of the 15th International Workshops on Conceptual Structures*, Sheffield, UK, 2007.
24. P.-F. Pai, Y.-J. Lyu and Y.-M. Wang, Analyzing academic achievement of junior high school students by an improved rough set model, *Computers & Education,* **54**(4), 2010, pp. 889–900.
25. P. D. Allison, *Missing Data*, Sage Publications, Thousand Oaks, CA, 2001.
26. R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, New York, NY, 1987.
27. J. L. Schafer, *Analysis of incomplete multivariate data*, London, 1997.
28. J. Bazan, H. Nguyen, S. Nguyen, P. Synak, J. Wroblewski, L. Polkowski, S. Tsumoto and T. Lin, Rough Set Algorithms in Classification Problem, in *Rough set methods and applications: new developments in knowledge discovery in information systems*, Physica-Verlag, 2000.
29. H. S. Nguyen, S. H. Nguyen and A. Skowron, Searching for features defined by hyperplanes, in *Foundations of Intelligent Systems*, Springer Berlin / Heidelberg, 1996.
30. S. H. Nguyen and H. S. Nguyen, Some efficient algorithms for rough set methods, in *Information Processing and Management of Uncertainty on Knowledge Based Systems*, Granada, Spain, 1996, pp. 1451–1456.
31. S. H. Nguyen, A. Skowron, P. Synak and J. Wroblewski, Knowledge discovery in databases: Rough set approach, in *Second Joint Annual Conference on Information Sciences*, Wrightsville Beach, North Carolina, 1997, pp. 34–37.
32. J. Bazan, A. Skowron, and P. Synak, Dynamic reducts as a tool for extracting laws from decisions tables, in *Methodologies for Intelligent Systems*, Springer Berlin / Heidelberg, 1994.
33. D. E. Goldberg, *GA in search, optimisation, and machine learning*, Addison-Wesley 1989.
34. J. H. Holland, *Adaptation in natural and artificial systems*, The MIT Press, Cambridge, 1992.
35. J. Bazan, Dynamic Reducts and Statistical Inference, in *Sixth International Conference on Information Procesing and Management of Uncertainty in Knowledge-Based Systems*, Granada, Spain, 1996, pp. 1147–1152.
36. G. Lee-Thomas, A. Kaw, and A. Yalcin, Using Online Endless Quizzes as Graded Homework, in *2011 Annual ASEE Conference and Exposition*, Vancouver, Canada, 2011.

Eleazar Gil-Herrera is a Doctoral Student in the Department of Industrial and Management System Engineering at the University of South Florida. In 2005, he received his B.S in Computer Engineering from the National University of Cusco, Peru. In 2009, Eleazar Gil-Herrera received his MS degree in Industrial Engineering from the University of Puerto Rico. His research interests are in the areas of knowledge discovery, data mining, decision support systems and prognostic models.

Athanasios Tsalatsanis is an Assistant Professor with the Center for Evidence-based Medicine, University of South Florida. In 2008, he received his Ph.D. in Industrial Engineering from the University of South Florida, in which he focused on the development of algorithms and control methodologies for autonomous robotic systems. He joined the Center for Evidence Based Medicine in 2009. Since then, his research has been concentrated on healthcare engineering, specifically in the areas of decision support systems, health information systems, prognostic models, and social network analysis.

Appendix B (continued)



Ali Yalcin is an Associate Professor of Industrial and Management Systems Engineering at the University of South Florida. He holds a Ph.D. in Industrial and Systems Engineering from Rutgers University. His research interests are in the areas of systems modeling and analysis, engineering education, information systems, and knowledge discovery. He has taught a variety of courses in Industrial Engineering in the areas of systems simulation, information systems, facilities design and linear systems. He also co-authored the 2006 Joint Publishers Book-of-the-Year textbook, Design of Industrial Information Systems, Elsevier.

Autar Kaw is a Professor of Mechanical Engineering and Jerome Krivanek Distinguished Teacher at the University of South Florida, USA. He holds a Ph.D. in Engineering Mechanics from Clemson University. His main scholarly interests are in engineering education, bascule bridge design, and mechanics of composite materials. With major funding from the US National Science Foundation, he is the lead developer of award-winning online resources for an undergraduate course in Numerical Methods (http://numericalmethods.eng.usf.edu). He is the recipient of the 2004 US Florida Professor of the Year Award from the Council for Advancement and Support of Education (CASE) and the Carnegie Foundation for the Advancement of Teaching (CFAT). He has authored several textbooks on subjects such as composite materials, numerical methods, matrix algebra, and computer programming.

Appendix C: Rough Set Theory Based Prognostication of Life Expectancy for Terminally Ill Patients[1]

[1] Copyright ©2011, IEEE. Reprinted, with permission, from Gil-Herrera, E.; Yalcin, A.; Tsatalsanis, A.; Barnes, L.E.; Djulbegovic, B., Rough Set Theory based prognostication of life expectancy for terminally ill patient, Engineering in Medicine and Biology Society, EMBS, 2011 Annual International Conference of the IEEE

33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Rough Set Theory based Prognostication of Life Expectancy for Terminally Ill Patients

Eleazar Gil-Herrera, Ali Yalcin, Athanasios Tsalatsanis, Laura E. Barnes and Benjamin Djulbegovic

Abstract—**We present a novel knowledge discovery methodology that relies on Rough Set Theory to predict the life expectancy of terminally ill patients in an effort to improve the hospice referral process. Life expectancy prognostication is particularly valuable for terminally ill patients since it enables them and their families to initiate end-of-life discussions and choose the most desired management strategy for the remainder of their lives. We utilize retrospective data from 9105 patients to demonstrate the design and implementation details of a series of classifiers developed to identify potential hospice candidates. Preliminary results confirm the efficacy of the proposed methodology. We envision our work as a part of a comprehensive decision support system designed to assist terminally ill patients in making end-of-life care decisions.**

I. Introduction

According to Medicare regulations, a patient should be referred to hospice if his/her life expectancy is less than 6 months [1]. However, despite the well-documented advantages of hospice services, terminally ill patients do not reap the maximum benefits of hospice care with the majority of them being referred to hospice either prematurely or too late. In general, premature hospice referral is translated to patients losing the opportunity to receive potentially effective treatment, which may have prolonged their lives. Conversely, late hospice referral reduces the quality of life for patients and their families. It is apparent that accurate prognostication of life expectancy is of vital importance for all parties involved in the hospice referral process (e.g. patients, their families, and their physicians).

Here, we propose a novel knowledge discovery methodology developed to identify terminally ill patients with life expectancy less than 6 months. The core of the proposed methodology is Rough Set Theory [2]. The rest of this paper describes implementation details, reports results, and discusses limitations and future directions of our work.

II. Methodology

A. Literature Review

Approaches for developing prognostic models for estimating survival for seriously ill patients range from the use of traditional statistical and probabilistic techniques [3]-[6], to models based on artificial intelligence techniques such as neural networks, decision trees and rough set methods [7]-[11]. A recent systematic review of prognostic tools for estimating survival in palliative care highlighted the lack of accurate end-of-life prognostic models [13].

Both statistics based techniques and *AI* based models rely on data that are precisely well defined. However, medical information, which represents patients records that include symptoms and clinical signs, is not always well defined and, therefore, the data are represented with vagueness [14]. Particularly, for this kind of information, it becomes very difficult to classify borderline cases in which very small differences in the value of a variable of interest may completely change categorization and therefore the following decisions can changes dramatically [15]. Moreover, the dataset is presented with inconsistencies in the sense that it is possible to have more than one patient with the same description but showing different outcomes.

In this work we propose the use of Rough Set Theory (RST) [2] to deal with vagueness and inconsistency in the representation of the dataset. RST provides a mathematical tool for representing and reasoning about vagueness and inconsistency. Its fundamentals are based on the construction of similarity relations between dataset objects from which approximate yet useful solutions are provided. In RST, the knowledge extracted from the data set is represented in the form of “if-then” decision rules where an explanation of how the final decision was derived can be traced. Clinical credibility in prognosis models depends on the ease with which practitioners and patients can understand and interpret the results [16]. Therefore, the if-then decision rule representation offers a significant advantage over “black box” modeling approaches such as neural networks.

RST has been used in a number of applications dealing with modeling medical prognosis [9]–[12]. For example, Tsumoto et al. [11], provides a framework to model medical diagnosis rules showing theoretically that the characteristics of medical reasoning reflect the concepts of approximation established in Rough Set Theory. Komorowski et al. [12], show that RST is useful to extract medical diagnosis rules to identify a group of patients for whom performing a test that is costly or invasive is redundant or superfluous in the prognosis of a particular medical condition.

In this paper we describe a RST based knowledge discovery methodology to provide a classifier that properly discriminates patients into two groups, those who survive at least 180 days after evaluation for hospice referral and those who do not. ROSETTA [17] software is used to perform the analysis described in the remainder of the paper.

Manuscript received March 26th, 2011. This work was supported in part by the Department of Army under grant #W81 XWH-09-2-0175.

E. Gil-Herrera and A. Yalcin are with the Department of Industrial and Management System Engineering, University of South Florida, Tampa, FL 33620, USA(e-mail: eleazar@mail.usf.edu, ayalcin@eng.usf.edu).

A. Tsalatsanis, L. E. Barnes and B. Djulbegovic are with the Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL 33612, USA (e-mail: atsalats@health.usf.edu, lbarnes@health.usf.edu, bdjulbeg@health.usf.edu).

U.S. Government work not protected by U.S. copyright

B. Dataset

The dataset used in this study consists of the 9105 cases from the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model dataset [18]. We consider all variables used in the SUPPORT prognostic model [4] as condition attributes, i.e. the physiologic variables along with the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function. Attributes' names and descriptions are listed in Table I.

As the decision attribute, we define a binary variable (Yes/No) "deceases_in_6months" using the following two attributes from the SUPPORT dataset:

TABLE I
CONDITION ATTRIBUTES

Name	Description
meanbp	Mean arterial blood pressure Day 3
wblc	White blood cell count Day 3
hrt	Heart rate Day 3
resp	Respiratory rate Day 3
temp	Temperature (Celsius)
alb	Serum Albumin
bili	Bilirubin
crea	Serum Creatinine
sod	Sodium
pafi	Pa02 / (.01 * FiO2)
ca	Presence of cancer
age	Patient's age
hday	Days in hospital at study admit
dzgroup	Diagnosis group
scoma	SUPPORT coma score based on Glasgow coma scale

- *"death"* which represents the event of death at any time up to NDI date (National Death Index date: Dec 31, 1994).
- *"D.time":* number of days of follow up

The values of the decision attribute are calculated converting the *"D.time"* value in months and comparing against the attribute *"death"* as follows:

- If "D.time" < 6 months and "death" is equal to 1 (the patient died within 6 months) then "deceases_in_6months" is equal to "Yes"
- If "D.time" > 6 months and "death" is equal to 1 (the patient died after 6 months) then "deceases_in_6months" is equal to "No"
- If "D.time" > 6 months and "death" is equal to 0 (the patient did not died after 6 months) then "deceases_in_6months" is equal to "No"

C. Rough Set Theory

Based on RST, we can formally define the prognostication problem as:

$$T = (U, A \cup \{d\}) \qquad (1)$$

where *T* represents the dataset in the form of a table. Each row represents an object and each column represents an attribute. *U* is a non-empty finite set of objects and the set *A* represents a non-empty finite set of attributes called the condition attributes. In our case, an object designates a terminally ill patient and an attribute designates each of the fifteen condition attributes that describe a patient (Table I). Also, for every attribute $a \in A$, the function a: $U \rightarrow V_a$ makes a correspondence between an object in *U* to an attribute value V_a which is called the value set of *a*.

The set *T* incorporates an additional attribute *{d}* called the decision attribute. The system represented by this scheme is called a *decision system*.

D. Rough Set Theory Based Knowledge Discovery Process

RST based knowledge discovery process requires sequential and parallel use of various mathematical, statistical and soft computing methodologies with the objective of identifying meaningful relationships between condition and decision attributes.

The selection of specific methodologies for knowledge discovery is largely dependent on the considered dataset. We have taken the following steps in our approach:

1) Data preprocessing: If the selected table contains "holes" in the form of missing values or empty cell entries; the table may be processed in various ways to yield a completed table in which all entries are present. The data completion process for SUPPORT dataset in [18] is adopted in this work. After the preprocessing phase, the number of patients with missing information is reduced by 2 cases. Therefore, there are 9103 complete cases.

The next step in preprocessing is the discretization process. 13 out of 15 of the conditional attributes are continuous; therefore we transformed them into categorical variables. The discretization process is based on the searching of cuts that determine intervals. This process enables the classifier in obtaining a higher quality of classification rules. We found that using cut-off defined by medical experts is the best alternative for the discretization process. We consider the APACHE III Scoring System [5] for determining the cut-off for the physiologic variables along with the age variable. The remaining variables, not defined in [5] are discretized using Boolean Reasoning Algorithm [19] implemented in the ROSETTA software.

Finally, the dataset is divided randomly into training and testing sets containing 500 and 8603 cases, respectively. The training set is used in the discretization process to obtain the cut-off for the numerical attributes.

2) Reduct Generation: This step reduces the dimensionality of the dataset with the intention of removing redundant information and consequently decreases the complexity of the mining process. A reduct is the minimal set of attributes that enable the same classification as the complete set of attributes without loss of information. There are many algorithms for computing reducts for which the effect to the classification performance is critical. Since the computational complexity of the reduct generation problem is NP-hard [19], various suboptimal techniques have been proposed. In this work the dynamic reduct approach ([20-21]) is used for reduct generation.

2.1) Dynamic Reducts

Dynamic reducts algorithm aims at obtaining the most stable sets of reducts for a given dataset by sampling within this dataset. Random samples of the testing set are selected iteratively and reducts for the samples are computed using genetic algorithms [22-23]. The reducts that most frequently appear in the samples are the most stable.

Based on the principle of the dynamic reducts technique, we have randomly selected 100 subdivisions of the training set to use for reduct generation. The actual number of patient profiles included in each subdivision of the training set varies between 50% and 90% of the training dataset. Using this approach, 229 reducts were obtained from which the set of decision rules are generated.

2.2) Using the decision attribute as condition attribute

Typically only the condition attributes are used to generate reducts. As an alternative, we included the decision attribute d in the set of condition attributes and calculated the reducts based on this scheme.

The decision attribute (deceases_in_6_months) used as a condition attribute is intended to represent the physician's estimate of life expectancy expressed in terms of the decision classes defined for this problem. Survival prognosis models that incorporate physician estimates are shown to improve both predictive accuracy and the ability to identify patients with high probabilities of survival or death [4]. In this case, 549 reducts were obtained. The next step is the induction of decision rules.

3) Rule Induction. The ultimate goal of the RST based knowledge discovery methodology is to generate decision rules, which will be used in classifying each patient as surviving or not surviving within the defined period of time. A decision rule has the form: *if A then B (A→ B),* where *A* is called the condition and *B* the decision of the rule. Decision rules can be thought of as a formal language for drawing conclusions from data.

The decision rules were generated based on the two aforementioned sets of reducts. After the process of reducts generation, the decision table is presented in a compact shape from which the decision rules are generated

4) Classification. Based on the set of rules generated, we can classify patients as surviving or not surviving the six-month period. However, not all rules are conclusive. Patients with profiles identical to the conditions of the rules are not decisively classified. In addition, there are situations of contradictory rules, e.g. one or more rules classify a patient as surviving and some other rules classify the same patient as dying. To overcome these problems a *standard voting* algorithm [19] is used which allows all rules to participate in the decision process and classify a patient based on majority voting.

III. RESULTS

This section compares the performance of the classification processes where, the patients in the training dataset are classified as *survive, not survive* or *undefined* based on the induced rules and the classification process described. The results are presented in a confusion matrix form.

The accuracy of each classification model is reported in terms of Area under the Receiver Operating Characteristic curve (AUC). The best possible classification is achieved when AUC is equal to 1, while no classification ability exists when AUC is equal to 0.5.

Table 2 presents the confusion matrix for the classification model based on reducts generated on only the original condition attributes (without including the decision attribute). Table 3 shows the confusion matrix for the alternative case where the decision attribute is included in the set of condition attributes.

TABLE 2

CONFUSION MATRIX. THE REDUCTS ARE BASED ON SET A. THE CLASSIFIER PRESENTS AUC EQUAL TO 0.55 INDICATING WEAK DISCRIMINATION ABILITY.

		Predicted		
		Not survive	Survive	Undefined
Actual	Not survive	1395	1953	677
	Survive	1410	2542	626

Sensitivity = 0.64
Specificity = 0.42
AUC = 0.55

TABLE 3

CONFUSION MATRIX. THE REDUCTS ARE BASED ON SET $A = A \cup \{d\}$. THE CLASSIFIER PRESENTS AUC EQUAL TO 0.90 INDICATING GOOD DISCRIMINATION ABILITY.

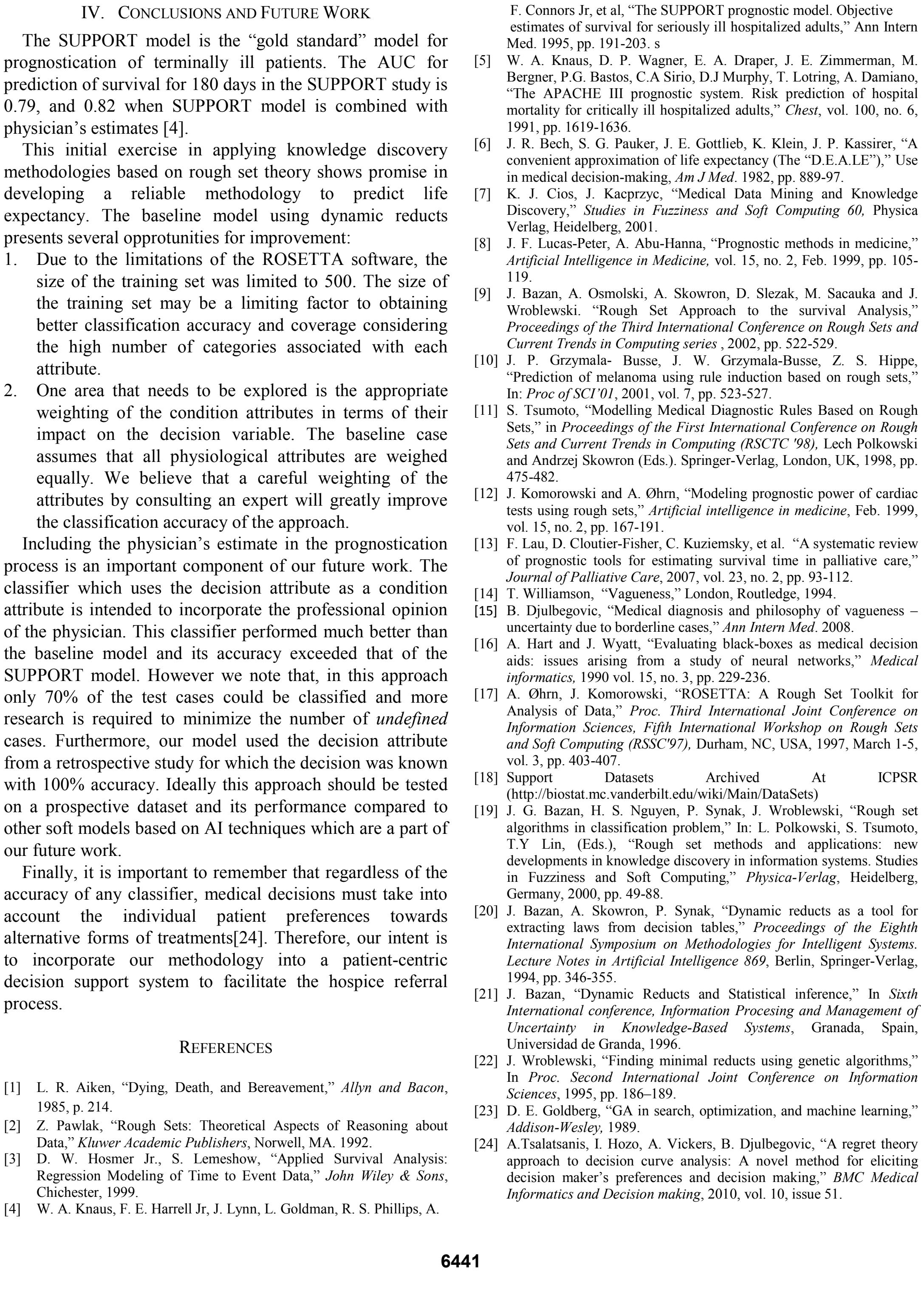
		Predicted		
		Not survive	Survive	Undefined
Actual	Not survive	1999	471	1555
	Survive	312	3245	1021

Sensitivity = 0.91
Specificity = 0.81
AUC = 0.90

The dynamic reducts approach without using the decision attribute as a condition attribute shows a weak discrimination ability. However, it demonstrates a fairly high level of coverage, being able to classify around 85% of the test cases. As shown in Table 3, the classification performance in terms of AUC when using the decision attribute as a part of the condition attributes is approximately 0.90. Both the specificity and sensitivity scores are tremendously improved. However, the classification coverage in this case is reduced to 70%.

The described classification process was repeated 10 times using randomly selected samples from the dataset (again 500 cases for training and the remainder 8603 cases for testing). The overall classification performance is obtaining by averaging the AUC from each iteration. Using the original set of attributes, the overall AUC is 0.56 (SD = 0.01). Following the same, we obtained an AUC of 0.85 (SD = 0.065) for the case where the decision attribute is used as a condition attribute.

Appendix C (continued)



IV. Conclusions and Future Work

The SUPPORT model is the "gold standard" model for prognostication of terminally ill patients. The AUC for prediction of survival for 180 days in the SUPPORT study is 0.79, and 0.82 when SUPPORT model is combined with physician's estimates [4].

This initial exercise in applying knowledge discovery methodologies based on rough set theory shows promise in developing a reliable methodology to predict life expectancy. The baseline model using dynamic reducts presents several opprotunities for improvement:

1. Due to the limitations of the ROSETTA software, the size of the training set was limited to 500. The size of the training set may be a limiting factor to obtaining better classification accuracy and coverage considering the high number of categories associated with each attribute.
2. One area that needs to be explored is the appropriate weighting of the condition attributes in terms of their impact on the decision variable. The baseline case assumes that all physiological attributes are weighed equally. We believe that a careful weighting of the attributes by consulting an expert will greatly improve the classification accuracy of the approach.

Including the physician's estimate in the prognostication process is an important component of our future work. The classifier which uses the decision attribute as a condition attribute is intended to incorporate the professional opinion of the physician. This classifier performed much better than the baseline model and its accuracy exceeded that of the SUPPORT model. However we note that, in this approach only 70% of the test cases could be classified and more research is required to minimize the number of *undefined* cases. Furthermore, our model used the decision attribute from a retrospective study for which the decision was known with 100% accuracy. Ideally this approach should be tested on a prospective dataset and its performance compared to other soft models based on AI techniques which are a part of our future work.

Finally, it is important to remember that regardless of the accuracy of any classifier, medical decisions must take into account the individual patient preferences towards alternative forms of treatments[24]. Therefore, our intent is to incorporate our methodology into a patient-centric decision support system to facilitate the hospice referral process.

References

[1] L. R. Aiken, "Dying, Death, and Bereavement," *Allyn and Bacon*, 1985, p. 214.

[2] Z. Pawlak, "Rough Sets: Theoretical Aspects of Reasoning about Data," *Kluwer Academic Publishers*, Norwell, MA. 1992.

[3] D. W. Hosmer Jr., S. Lemeshow, "Applied Survival Analysis: Regression Modeling of Time to Event Data," *John Wiley & Sons*, Chichester, 1999.

[4] W. A. Knaus, F. E. Harrell Jr, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors Jr, et al, "The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults," Ann Intern Med. 1995, pp. 191-203. s

[5] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P.G. Bastos, C.A Sirio, D.J Murphy, T. Lotring, A. Damiano, "The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults," *Chest*, vol. 100, no. 6, 1991, pp. 1619-1636.

[6] J. R. Bech, S. G. Pauker, J. E. Gottlieb, K. Klein, J. P. Kassirer, "A convenient approximation of life expectancy (The "D.E.A.LE")," Use in medical decision-making, *Am J Med*. 1982, pp. 889-97.

[7] K. J. Cios, J. Kacprzyc, "Medical Data Mining and Knowledge Discovery," *Studies in Fuzziness and Soft Computing 60,* Physica Verlag, Heidelberg, 2001.

[8] J. F. Lucas-Peter, A. Abu-Hanna, "Prognostic methods in medicine," *Artificial Intelligence in Medicine,* vol. 15, no. 2, Feb. 1999, pp. 105-119.

[9] J. Bazan, A. Osmolski, A. Skowron, D. Slezak, M. Sacauka and J. Wroblewski. "Rough Set Approach to the survival Analysis," *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing series* , 2002, pp. 522-529.

[10] J. P. Grzymala- Busse, J. W. Grzymala-Busse, Z. S. Hippe, "Prediction of melanoma using rule induction based on rough sets," In: *Proc of SCI'01*, 2001, vol. 7, pp. 523-527.

[11] S. Tsumoto, "Modelling Medical Diagnostic Rules Based on Rough Sets," in *Proceedings of the First International Conference on Rough Sets and Current Trends in Computing (RSCTC '98),* Lech Polkowski and Andrzej Skowron (Eds.). Springer-Verlag, London, UK, 1998, pp. 475-482.

[12] J. Komorowski and A. Øhrn, "Modeling prognostic power of cardiac tests using rough sets," *Artificial intelligence in medicine*, Feb. 1999, vol. 15, no. 2, pp. 167-191.

[13] F. Lau, D. Cloutier-Fisher, C. Kuziemsky, et al. "A systematic review of prognostic tools for estimating survival time in palliative care," *Journal of Palliative Care*, 2007, vol. 23, no. 2, pp. 93-112.

[14] T. Williamson, "Vagueness," London, Routledge, 1994.

[15] B. Djulbegovic, "Medical diagnosis and philosophy of vagueness – uncertainty due to borderline cases," *Ann Intern Med*. 2008.

[16] A. Hart and J. Wyatt, "Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks," *Medical informatics,* 1990 vol. 15, no. 3, pp. 229-236.

[17] A. Øhrn, J. Komorowski, "ROSETTA: A Rough Set Toolkit for Analysis of Data," *Proc. Third International Joint Conference on Information Sciences, Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97),* Durham, NC, USA, 1997, March 1-5, vol. 3, pp. 403-407.

[18] Support Datasets Archived At ICPSR (http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets)

[19] J. G. Bazan, H. S. Nguyen, P. Synak, J. Wroblewski, "Rough set algorithms in classification problem," In: L. Polkowski, S. Tsumoto, T.Y Lin, (Eds.), "Rough set methods and applications: new developments in knowledge discovery in information systems. Studies in Fuzziness and Soft Computing," *Physica-Verlag*, Heidelberg, Germany, 2000, pp. 49-88.

[20] J. Bazan, A. Skowron, P. Synak, "Dynamic reducts as a tool for extracting laws from decision tables," *Proceedings of the Eighth International Symposium on Methodologies for Intelligent Systems. Lecture Notes in Artificial Intelligence 869*, Berlin, Springer-Verlag, 1994, pp. 346-355.

[21] J. Bazan, "Dynamic Reducts and Statistical inference," In *Sixth International conference, Information Procesing and Management of Uncertainty in Knowledge-Based Systems*, Granada, Spain, Universidad de Granda, 1996.

[22] J. Wroblewski, "Finding minimal reducts using genetic algorithms," In *Proc. Second International Joint Conference on Information Sciences*, 1995, pp. 186–189.

[23] D. E. Goldberg, "GA in search, optimization, and machine learning," *Addison-Wesley,* 1989.

[24] A.Tsalatsanis, I. Hozo, A. Vickers, B. Djulbegovic, "A regret theory approach to decision curve analysis: A novel method for eliciting decision maker's preferences and decision making," *BMC Medical Informatics and Decision making*, 2010, vol. 10, issue 51.

Appendix D: Towards a Classification Model to Identify Hospice Candidates in Terminally Ill Patients[1]

[1] Copyright ©2012, IEEE. Reprinted, with permission, from Gil-Herrera, E.; Yalcin, A.; Tsatalsanis, A.; Barnes, L.E.; Djulbegovic, B., Towards a Classification Model to Identify Hospice Candidates in Terminally Ill Patients, Engineering in Medicine and Biology Society, EMBS, 2012 Annual International Conference of the IEEE

Appendix D (continued)

34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Towards a Classification Model to Identify Hospice Candidates in Terminally Ill Patients

Eleazar Gil-Herrera, Ali Yalcin, Athanasios Tsalatsanis, Laura E. Barnes and Benjamin Djulbegovic

Abstract— **This paper presents a Rough Set Theory (RST) based classification model to identify hospice candidates within a group of terminally ill patients. Hospice care considerations are particularly valuable for terminally ill patients since they enable patients and their families to initiate end-of-life discussions and choose the most desired management strategy for the remainder of their lives. Unlike traditional data mining methodologies, our approach seeks to identify subgroups of patients possessing common characteristics that distinguish them from other subgroups in the dataset. Thus, heterogeneity in the data set is captured before the classification model is built. Object related reducts are used to obtain the minimum set of attributes that describe each subgroup existing in the dataset. As a result, a collection of decision rules is derived for classifying new patients based on the subgroup to which they belong. Results show improvements in the classification accuracy compared to a traditional RST methodology, in which patient diversity is not considered. We envision our work as a part of a comprehensive decision support system designed to facilitate end-of-life care decisions. Retrospective data from 9105 patients is used to demonstrate the design and implementation details of the classification model.**

I. Introduction

A. Hospice referral criteria

Hospice is designed to provide comfort and support to terminally ill patients and their families. According to Medicare regulations, a patient should be referred to hospice if his/her life expectancy is approximately 6 months or less [1]. However, most patients are not referred to hospice in a timely manner [2, 3] and therefore they do not reap the well-documented benefits of hospice services. A premature hospice referral translates to a patient losing the opportunity to receive potentially effective treatment, which may prolong their life. Conversely, a late hospice referral may deprive patients and their families of enjoying the benefits offered. Therefore, accurate prognostication of life expectancy is of vital importance for terminal patients as well as for their families and physicians.

B. Prognostic models for estimating survival of terminally ill patients

Survival prognostic models range from traditional statistical and probabilistic techniques [4-10], to models based on artificial intelligence such as neural networks [11, 12], decision trees [13, 14] and rough set methods [15, 16]. The primary goal of survival prognostic models is to provide accurate information regarding life expectancy and/or determine the association between prognostic factors and survival. Typically, the information derived by prognostic models is presented in terms of probability of death within a time period. Recent systematic reviews [17, 18] have highlighted the necessity of prediction models that can be easily integrated into clinical practice and facilitate end-of-life clinical decision-making.

Several important issues demand particular consideration when developing clinical classification models: First, clinical data, representing patient records that include symptoms and clinical signs, are not always well defined and are represented with *vagueness* [19]. Therefore, it is very difficult to classify cases in which small differences in the value of an attribute may completely change the classification of a patient and, as a result, the treatment decisions [20]. Second, clinical data may present *inconsistencies*, which means that it is possible to have more than one patient with the same description but with different outcomes. Third, the results of prognostic models should be readily interpretable to enable practical and posteriori inspection and interpretation by the treating physician or an expert system [21]. Finally, prognostic models should consider the heterogeneity in clinical data, i.e. the existence of patient diversity presented in terms of risk of disease and responsiveness to treatment [22, 23]. This consideration will enable a prognostic model to identify possible subgroups of patients for which certain covariates do not influence their classification. The practical implications of such considerations are associated with the ability to customize the prognostic model for each subgroup of patients (e.g. expensive and/or potentially harmful tests may be avoided for particular subgroups).

Rough Set Theory (RST) [24], a mathematical tool for representing and reasoning about vagueness and inconsistency in data sets, has been used in a number of applications dealing with modeling medical prognosis [15, 16, 25-28]. For example, Tsumoto et al. [25], provide a framework to model medical diagnosis rules showing theoretically that the characteristics of medical reasoning reflect the concepts of approximation established in RST. Komorowski et al. [26], show that RST is useful to extract medical diagnosis rules to identify a group of patients for whom performing a test that is costly or invasive is redundant or superfluous in the prognosis of a particular medical condition. Recently, [28] highlighted features of RST for integrating into medical applications. For example, RST has the ability to handle imprecise and uncertain information and provides a schematic approach for analyzing data without initial assumptions on data distribution.

This work was supported in part by the Department of Army under grant #W81 XWH-09-2-0175.

E. Gil-Herrera and A. Yalcin are with the Department of Industrial and Management System Engineering, University of South Florida, Tampa, FL 33620, USA(e-mail: eleazar@mail.usf.edu, ayalcin@eng.usf.edu).

A. Tsalatsanis, L. Barnes and B. Djulbegovic are with the Center for Evidence Based Medicine and Health Outcomes Research, University of South Florida, Tampa, FL 33612, USA (e-mail: atsalats@health.usf.edu, lbarnes@health.usf.edu, bdjulbeg@health.usf.edu)

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

In our previous work [29], we proposed the use of RST to predict the life expectancy of terminally ill patients using a *global reduction* [30] methodology to identify the most significant attributes for building the classification model. However, we found that the number of attributes used in the model was barely reduced and therefore produced long decision rules. Moreover, considering the number of discretization categories associated with each attribute, the generated decision rules were built to describe each object in the training set and therefore, they were poorly suited for classifying new cases.

Here, we propose the use of an alternative attribute reduction methodology that aims to identify groups of patients that share common characteristics that distinguish them from the rest of the patients. As a result, we obtain subgroups of patients from which different sets of significant attributes are identified. The decision rules generated in this manner contain fewer attributes and therefore are more suitable to classify new patients. Moreover, by studying each subgroup, we can reason about how a different rule-set is applied to a particular patient.

The rest of the paper describes details of the proposed RST based methodology to provide a classifier that properly discriminates patients into two groups: those who survive at least 180 days after evaluation for hospice referral and those who do not. ROSETTA [31] software is used to perform the analysis described in the remainder of the paper.

II. Methodology

A. Data Set

The dataset used in this study consists of the 9105 cases from the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model dataset [30]. We consider all variables used in the SUPPORT prognostic model [3] as condition attributes, i.e. the 10 physiologic variables along with the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function. Data collection and patient selection procedures are detailed in [3]. Attributes names and descriptions are listed in Table I. As the decision attribute, we define a binary variable (Yes/No) "deceases_in_6months" using the following two attributes from the SUPPORT prognosis model dataset:

- death: represents the event of death at any time up to NDI date (National Death Index date: Dec 31, 1994).
- D.time: number of days of follow up

The values of the decision attribute are calculated converting the "D.time" value in months and comparing against the attribute "death" as follows:

- If "D.time" < 6 months and "death" is equal to 1 (the patient died within 6 months) then "deceased_in_6months" is "Yes". Otherwise, it is implicit that a patient survived the 6-month period; hence, "deceased_in_6months" is "No".

B. Rough Set Theory Data Representation

Based on RST, the data set is represented as:

$$T = (U, A \cup \{d\}) \quad (1)$$

TABLE I. Condition Attributes

Name	Description
alb	Serum albumin
bili	Bilirubin
crea	Serum creatinine
hrt	Heart rate
meanbp	Mean arterial blood pressure
pafi	Arterial blood gases
resp	Respiratory rate
sod	Sodium
temp	Temperature (Celsius)
wblc	White blood cell count
dzgroup	Diagnosis group
age	Patient's age
hday	Days in hospital at study admit
ca	Presence of cancer
scoma	SUPPORT coma score based on Glasgow coma scale

where *T,* represents the dataset in the form of a table. Each row represents an object and each column represents an attribute. *U* is a non-empty finite set of objects and the set *A* is a non-empty finite set of attributes called the condition attributes. In our case, an object designates a terminally ill patient and an attribute $a \in A$ designates each of the fifteen condition attributes that describe a patient (Table I). For every attribute, the function $a: U \rightarrow V_a$ makes a correspondence between an object in *U* to an attribute value V_a which is called the value set of a. The set T incorporates an additional attribute *{d}* called the decision attribute. The system represented by this scheme is called a decision system.

C. Development of the Classification Model

This process typically involves numerous steps, such as data preprocessing, discretization, reduction of attributes, rule induction, classification and interpretation of the results. Details on the data preprocessing and data discretization for this data set are described in [29]. The ultimate goal of this process is to generate decision rules, which are used to classify each patient as surviving or not surviving within the defined period of time. A decision rule has the form: *if A then B (A → B),* where *A* is called the condition and *B* the decision of the rule.

Here, we are focusing on an alternative method of reducing the attribute dimensions and identify different subgroups of similar patients in the data set. In [32], two types of reducts are defined:

1) Global Reducts:

Consists of the minimal set of attributes that preserve the structure of the entire data set. A set $B \subseteq A$ is called a global reduct if the indiscernibility relation using attributes B is equal to the indiscernibility relation using all the condition attributes A, i.e.:

$$IND(B) = IND(A), \text{ where,}$$

$$IND(B) = \{(u_i, u_j) \in U^2 \colon \forall\, a_k \in B, \mathrm{a_k(u_i)} \neq \mathrm{a_k(u_j)}\}$$

As an example, consider the following global reduct obtained from the data set containing 12 condition attributes:

G_RED = {age, dzgroup, scoma, ca, meanbp, wblc, hrt, resp, temp, bili, crea, sod}

Using *G_RED*, few patients will have exactly the same attribute-value combinations because the number of discretization categories associated with each attribute is high. Thus, the decision rules generated are too specific to the cases in the training set and therefore may not be able to classify new cases accurately. Moreover, the fact that global reducts represent the entire data set makes it difficult to detect the presence of heterogeneous groups in the data meaning that the causes of diversity between the patient outcomes will remain unknown.

2) *Object related reducts (ORR):*

Represents the minimal attribute subsets that discern an object $u \in U$ from the rest of objects belonging to a different decision class. Mathematically, an ORR $R_u \subseteq A$ is defined as:

$$\forall u_i \in U : d(u_i) \neq d(u_j) \Rightarrow \exists a_k \in R_u : a_k(u_i) \neq a_k(u_j), \; where\ u_i \neq u_j .$$

An ORR is the minimal and vital information that is used to partition the universe of objects into smaller, homogeneous subgroups, where objects within a subgroup are related by means of information described by the ORR. Decision rules generated by this scheme will usually contain fewer attributes and are more suitable to classify new cases. Some decision rules contain a different set of attributes applicable for a particular subgroup of patients.

III. Results

The two methods for dimensionality reduction produce a set of reducts. The number of reducts and decision rules obtained are presented in Table II. Based on the decision rules generated, patients are classified as surviving or not surviving the six-month period. A standard voting algorithm [30] is used for this purpose. Table III, presents the performance of two classification models based on each type of reduct generation described. The performance of each classification model is represented in terms of *sensitivity*, *specificity*, *Area under the Receiver Operating Characteristic curve* (AUC) and *coverage* of the model. A 5-fold cross validation procedure was applied to estimate the performance of each classification model, where, the entire data set is randomly divided into five subsets (folds). Then, each fold (20% of the data set) is used once as a testing set, while the remaining folds (80%) are used for training. The process is repeated five times and the results are averaged to provide an estimate for the classifier performance.

Compared to the Global reduct approach, the ORR approach has enhanced the classification performance in terms of AUC and sensitivity. Moreover the decision rules generated are able to classify all new cases.

IV. Discussion

Analyzing the information obtained from the ORR, we can identify groups of patients for whom it is possible to evade costly, invasive or even unnecessary tests required by the prediction model. For example, the following two ORRs generate rules independent of the *Pafi* score (associated with the patient's blood gases), without reducing the classification accuracy. The importance of such finding becomes apparent considering that in clinical practice *Pafi* is not collected routinely for patients outside the Intensive Care Unit (ICU).

TABLE II. Number of reducts and decision rules generated – Global vs. ORR

Method	Number of reducts	Number of rules
Global reducts	99	647,223
ORR	11,894	68,492

TABLE III. Clasification results – Global vs. ORR

Method	Sensitivity	Specificity	AUC	Coverage
Global reducts	73.67%	44.05%	61.8%	86.43%
ORR	86.92%	39.2%	71.9%	100%

- ORR = {Age, dzgroup, meanbp} generates the following decision rules:
  - if age= [45, 60) AND dzgroup = (Lung Cancer) AND meanbp=[60, 70) then: Survive = 22.86%, Die = 77.14%.
  - if age= [45, 60) AND dzgroup = (CHF) AND meanbp=[100, 120) then: Survive = 82.93%, Die = 17.07%.
  - if age= [70, 75) AND dzgroup = (COPD) AND meanbp=[80,100) then: Survive = 84.21%, Die = 15.79%.
- ORR = {Age, dzgroup, hrt, crea} generates the following decision rules:
  - if age= [45, 60) AND dzgroup = (CHF) AND hrt=[100,110) and crea[1.95, *] then: Survive = 83.33%, Die = 16.67%.
  - if age= [75,85) AND dzgroup = (CHF) AND hrt=[50,110) and crea[0.5, 1.5) then: Survive = 82.19%, Die = 17.81%.

Consequently, the use of *Pafi* test in patients that belong to one of those groups defined by the ORR's will not improve the prognostication accuracy.

Our approach demonstrates features that make it particularly suitable for use in clinical decision-making. It is a patient-centric methodology which is able to predict without the use of unnecessary, expensive and/or invasive procedures for certain subgroups of patients. Consequently, selection of attributes upon which a decision is to be made is critical to minimizing healthcare costs and maximizing the quality of patient care. Finally, considering that more than one ORR could discern each patient, the information acquired offers several options dependent on the attribute values available for each individual patient.

V. Future Work

The number of ORR and the decision rules generated depends on the number of condition attributes and its categories. For clinical datasets, which contain large numbers of condition attributes, the number of ORRs and decision rules generated can be extremely large to be evaluated directly by human experts. Therefore, the interpretation and analysis of the ORRs and their decision rules requires the use of a well-defined methodology.

Compared to our previous work [29], the results presented in this paper show an improvement in the classifier performance. However, further research need to be conducted in order to achieve a reliable prognostic model.

References

[1] L.R. Aiken and NetLibrary Inc., “Dying, death, and bereavement,” in Book Dying, death, and bereavement, *Series Dying, death, and bereavement*, 4th ed. Lawrence Erlbaum Associates, 2000.

[2] N.A. Christakis, “Timing of referral of terminally ill patients to an outpatient hospice.,” *J Gen Intern Med*, vol. 9, (no. 6), pp. 314-20, Jun 1994.

[3] A. Tsalatsanis, L.E. Barnes, I. Hozo, and B. Djulbegovic, “Extensions to Regret-based Decision Curve Analysis: An application to hospice referral for terminal patients,” *BMC Med Inform Decis Mak*, vol. 11, pp. 77, 2011.

[4] W.A. Knaus, D.P. Wagner, E.A. Draper, J.E. Zimmerman, M. Bergner, P.G. Bastos, C.A. Sirio, D.J. Murphy, T. Lotring, and A. Damiano, “The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults,” *Chest*, vol. 100, (no. 6), pp. 1619-1636, December, 1991.

[5] W.A. Knaus, F.E. Harrell, J. Lynn, L. Goldman, R.S. Phillips, A.F. Connors, N.V. Dawson, W.J. Fulkerson, R.M. Califf, N. Desbiens, P. Layde, R.K. Oye, P.E. Bellamy, R.B. Hakim, and D.P. Wagner, “The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults,” *Annals of Internal Medicine*, vol. 122, (no. 3), pp. 191-203, February, 1995.

[6] D.W. Hosmer and S. Lemeshow, *Applied survival analysis regression modeling of time to event data*, New York, NY: Wiley, 1999.

[7] J.R. Beck, S.G. Pauker, J.E. Gottlieb, K. Klein, and J.P. Kassirer, “A convenient approximation of life expectancy (the “DEALE”): II. Use in medical decision-making,” *The American Journal of Medicine*, vol. 73, (no. 6), pp. 889-897, 1982.

[8] I. Hyodo, T. Morita, I. Adachi, Y. Shima, A. Yoshizawa, and K. Hiraga, “Development of a Predicting Tool for Survival of Terminally Ill Cancer Patients,” *Japanese Journal of Clinical Oncology*, vol. 40, (no. 5), pp. 442-448, May 1, 2010.

[9] D. Porock, D. Parker-Oliver, G. Petroski, and M. Rantz, “The MDS Mortality Risk Index: The evolution of a method for predicting 6-month mortality in nursing home residents,” *BMC Research Notes*, vol. 3, (no. 1), pp. 200, 2010.

[10] P.K.J. Han, M. Lee, B.B. Reeve, A.B. Mariotto, Z. Wang, R.D. Hays, K.R. Yabroff, M. Topor, and E.J. Feuer, “Development of a Prognostic Model for Six-Month Mortality in Older Adults With Declining Health,” *Journal of Pain and Symptom Management*, vol. 43, (no. 3), pp. 527-539, 2012.

[11] G. Clermont, D.C. Angus, S.M. DiRusso, M. Griffin, and W.T. Linde-Zwirble, “Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models,” *Critical Care Medicine*, vol. 29, (no. 2), 2001.

[12] R. Dybowski, V. Gant, P. Weller, and R. Chang, “Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm,” *The Lancet*, vol. 347, (no. 9009), pp. 1146-1150, 1996.

[13] M.R. Segal, “Features of Tree-Structured Survival Analysis,” *Epidemiology*, vol. 8, (no. 4), pp. 344-346, 1997.

[14] S.S. Hwang, C.B. Scott, V.T. Chang, J. Cogswell, S. Srinivas, and B. Kasimis, “Prediction of Survival for Advanced Cancer Patients by Recursive Partitioning Analysis: Role of Karnofsky Performance Status, Quality of Life, and Symptom Distress,” *Cancer Investigation*, vol. 22, (no. 5), pp. 678-687,2004.

[15] J. Bazan, A. Osmólski, A. Skowron, D. Ślçezak, M. Szczuka, and J. Wróblewski, “Rough Set Approach to the Survival Analysis - Rough Sets and Current Trends in Computing,” vol. 2475, *Lecture Notes in Computer Science*, J. Alpigini, J. Peters, A. Skowron and N. Zhong eds.: Springer Berlin / Heidelberg, pp. 951-951, 2002.

[16] P. Pattaraintakorn, N. Cercone, and K. Naruedomkul, “Hybrid rough sets intelligent system architecture for survival analysis,” in Transactions on rough sets VII, W. M. Victor, O. Ewa, owska, S. Roman, owinski and Z. Wojciech eds.: Springer-Verlag, 2007, pp. 206-224.

[17] F. Lau, D. Cloutier-Fisher, C. Kuziemsky, F. Black, M. Downing, and E. Borycki, *A systematic review of prognostic tools for estimating survival time in palliative care*, Montreal, CANADA: Centre of Bioethics, Clinical Research Institute of Montreal, 2007.

[18] P. Glare, C. Sinclair, M. Downing, P. Stone, M. Maltoni, and A. Vigano, “Predicting survival in patients with advanced disease,” *European Journal of Cancer*, vol. 44, (no. 8), pp. 1146-1156, 2008.

[19] P. Simons, “VAGUENESS” *International Journal of Philosophical Studies*, vol. 4, (no. 2), pp. 321-327, Sep 1996.

[20] B. Djulbegovic, “Medical diagnosis and philosophy of vagueness-uncertainty due to borderline cases,” *Annals of Internal Medicine*, 2008.

[21] J.C. Wyatt and D.G. Altman, “Commentary: Prognostic models: clinically useful or quickly forgotten?,” *BMJ*, vol. 311, (no. 7019), pp. 1539-1541, 1995.

[22] R.L. Kravitz, N. Duan, and J. Braslow, “Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages,” *Milbank Quarterly*, vol. 82, (no. 4), pp. 661-687, 2004.

[23] P. Schlattmann, “Introduction - Heterogeneity in Medicine Medical Applications of Finite Mixture Models,” *Statistics for Biology and Health*: Springer Berlin Heidelberg, 2009, pp. 1-22.

[24] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Norwell, MA, 1992.

[25] S. Tsumoto, “Modelling Medical Diagnostic Rules Based on Rough Sets- Rough Sets and Current Trends in Computing,” vol. 1424, *Lecture Notes in Computer Science*, L. Polkowski and A. Skowron eds.: Springer Berlin / Heidelberg, 1998, pp. 475-482.

[26] J. Komorowski and A. Øhrn, “Modelling prognostic power of cardiac tests using rough sets,” *Artificial Intelligence in Medicine*, vol. 15, (no. 2), pp. 167-191, 1999.

[27] P. Grzymala-Busse, J.W. Grzymala-Busse, and Z.S. Hippe, “Melanoma prediction using data mining system LERS,” in *Proc,, COMPSAC,* 2001, pp. 615-620.

[28] P. Pattaraintakorn and N. Cercone, “Integrating rough set theory and medical applications,” *Applied Mathematics Letters*, vol. 21, (no. 4), pp. 400-403, 2008.

[29] E. Gil-Herrera, A. Yalcin, A. Tsalatsanis, L.E. Barnes, and D. B, “Rough Set Theory based Prognostication of Life Expectancy for Terminally Ill Patients,” in *Conf Proc IEEE Eng Med Biol Soc*, 2011, pp. 6438-6441.

[30] J. Bazan, H. Nguyen, S. Nguyen, P. Synak, J. Wroblewski, L. Polkowski, S. Tsumoto, and T. Lin, “Rough Set Algorithms in Classification Problem,” in Rough set methods and applications: new developments in knowledge discovery in information systems: Physica-Verlag, 2000, pp. 49-88.

[31] Ø. Alexander and J. Komorowski, “ROSETTA: A Rough Set Toolkit for Analysis of Data,” in *Proc. Third International Joint Conference on Information Sciences*, 1997, pp. 403-407.

[32] J.G. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak and J. Wróblewski, “Rough set algorithms in classification problem,” in *Rough set methods and applications*: Physica-Verlag GmbH, 2000, pp. 49-88.

Appendix E: Rough Set Theory Based Prognostic Models for Hospice Referral [1]

[1] In second round of review in the Artificial Intelligence in Medicine Journal

Appendix E (continued)

Rough Set Theory Based Prognostic Models for Hospice Referral

Eleazar Gil-Herrera[a,*], Garrick Aden-Buie[a], Ali Yalcin[a], Athanasios Tsalatsanis[b], Laura E. Barnes[c], Benjamin Djulbegovic[d,e]

[a] *Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL 33620, USA*
[b] *Health Informatics and Decision Making, Clinical & Translational Science Institute, Department of Internal Medicine, University of South Florida, Tampa, FL 33612, USA*
[c] *Department of Systems and Information Engineering, University of Virginia, 151 Engineer's Way, Charlottesville, VA 22904, USA*
[d] *Center for Evidence-based Medicine and Health Outcomes Research, Department of Internal Medicine, 12901 Bruce B Downs Blvd. MDC 27, Tampa, FL, 33612 USA*
[e] *H. Lee Moffitt Cancer Center & Research Institute, Departments of Hematology and Health Outcomes and Behavior, 12902 Magnolia Drive, Tampa, FL 33612, USA*

Abstract

Objective: The goal of this paper is to explore and evaluate the application of classical and dominance-based Rough Set Theory (RST) for the development of data-driven prognostic models for hospice referral. In this work, rough set based prognostic models are compared with other data-driven methods with respect to two factors related to clinical credibility: accuracy and accessibility.

Methods: We utilize retrospective data from 9,103 terminally ill patients to demonstrate the design and implementation of classical and dominance-based RST classification models to identify potential hospice candidates. The classical rough set approach (CRSA) provides methods for knowledge acquisition, founded on the relational indiscernibility of objects in a decision system, to describe required conditions for membership in a concept class. On the other hand, the dominance-based rough set approach (DRSA) analyzes information based on the monotonic relationships between condition attributes values and their assignment to the decision class. CRSA decision rules for six-month patient survival classification were induced under the classical rough set approach using the MODLEM algorithm. Dominance-based decision rules were extracted from the dataset utilizing the VC-DomLEM rule induction algorithm.

Results: The RSA classifiers are compared with other predictive and rule based decision modeling techniques by examining the accuracy and accessibility of the models. Accessible prognostic models provide traceable, interpretable results and use reliable data. Both classical and dominance-based RSA classifiers perform comparably in accuracy to other common classification methods, while providing significant advantages in terms of traceability and interpretability of the model.

Conclusions: This paper contributes to the growing body of research in RST-based prognostic models. RST and its extensions posses features that enhance the accessibility of clinical decision support models. Developing prognostic models for hospice referrals is a challenging problem resulting in substandard performance for all of the evaluated classification methods.

Keywords: rough set theory, dominance-based rough set approach, hospice referral, prognostic models

Preprint submitted to Elsevier *October 18, 2013*

1. Introduction

Hospice care reduces the emotional burden of illness on terminal patients by optimizing pain relief strategies [1] and provides a demonstrated, cost-effective increase in the quality of end-of-life care when compared to conventional programs [2]. This increase in quality of care elevates the quality of life of both patients and their families [3].

The advantages of hospice care are diminished for terminally ill patients who enter either prematurely or too late. In general, premature hospice referral represents a lost opportunity for the patient to receive potentially effective and life-prolonging treatment. Conversely, late hospice referral is not desirable and negatively impacts both the quality of end-of-life care and the quality of life of patients and their families [4, 5]. According to Medicare regulations, patient eligibility for hospice care is contingent upon a life expectancy of less than six months, as estimated by the attending physician and certified by the medical director of the hospice program [6]. Medicare claims data report that 14.9% of hospice care patients lived for more than 180 days after enrollment, while 28.5% were late referrals who died within 14 days [4, 6]. Accurate prognostication of life expectancy is crucial in end-of-life care decisions and is consequently of vital importance for patients, their physicians and their families.

Prognostic models are an important instrument in prognostication as, in conjunction with direct physician observation, they increase the accuracy of prognostication when compared to physician observation alone [7]. However, a significant barrier to the widespread practical use of prognostic models is their perceived lack of clinical credibility [8].

The objective of this work is to explore and evaluate the application of rough set approaches in the development of data-driven prognostic models with respect to two criteria essential to clinical credibilty: accuracy and accessibility. To this end, we will explore Rough Set Theory as it is applied to end-of-life care and hospice referral decision support models.

This paper is organized as follows: in Section 2 we present important features of clinically credibile prognostic models and other characteristics of clinical data sets that motivate the use of Rough Set Theory (RST). In Section 3.1, we present an overview of the fundamental theory of rough sets for analyzing datasets, and in Section 3.2 we present a similar overview of the theory of the Dominance-based Rough Set Approach (DRSA). In Section 3.3 we discuss the use of decision rules in conjunction with the rough set approaches. Section 3.4 describes the dataset used for the demonstration of the proposed prognostic models. Section 3.5 presents the development of the prognostic models, followed, in Section 3.6, by an overview of the performance evaluation methods used in this study. Finally, Sections 4, 5, and 6 report results and conclusions, and discuss limitations and future directions of our work.

2. Motivation

The objective of a prognostic model is to determine quantitative or symbolic relationships between covariates and a health-related outcome. In the case of life expectancy estimation, prognostic models improve the accuracy in critical clinical decisions and are shown to be superior to physicians' prognostication alone [9]. Models for estimating the life expectancy of terminally ill patients include the use of statistical and probabilistic methods [10–18], artificial intelligence techniques such as neural networks and support vector machines (SVM) [19–21], decision

[☆] The authors declare no conflicts of interest.

[*] Corresponding author. Phone: (813) 974-9453; email: `eleazar@mail.usf.edu`

trees [22, 23] and rough set methods [24, 25]. Survival models [6, 12, 14, 16, 18, 22, 23] focus on estimating the probability that a patient will survive a finite period of time. Classification models, based on methods such as neural networks, SVM and logistic regression [17, 19–21, 26], represent the survival outcome as a binary variable, predicting the status of a patient at a critical point in time (e.g. six months) by classifying the patient as surviving or not surviving the critical time frame.

A recent review [15] demonstrated that, despite the importance of accurate prognostication within the spectrum of medical care objectives, there is a lack of accessible and accurate prognostic models available to physicians in practice. To withstand clinical trials, and to meet the needs of physicians and patients, a prognostic model must have clinical credibility, meaning that the model must posses a high level of accuracy and accessibility for physicians to believe in the value of the model as a prognostic tool. That is, in addition to accurate prognostication, such a model should be traceable in its structure, meaning the “model’s structure should be apparent and its predictions should make sense to the doctors who will rely on them” [8]. Likewise, the model should provide interpretable results that facilitate explanation of the prognosis, the data required for the model must be relevant and simple to collect with high reliability, and physicians must be able to apply the modeling method correctly without violating the fundamental assumptions of the model.

Clinical datasets present unique challenges that must also be addressed when building data-driven prognostic models. Cios and Moore [27] argue that there are a number of features specific to medical data that result from the volume, heterogeneity and complexity of data that lack canonical form and are limited by significant ethical, legal and social constraints. Furthermore, the underlying conceptual structures of medicine are not easily formalized mathematically, as the medical field lacks the necessary constraints for the mathematical characterizations common to the physical sciences. As a result, many medical concepts are *vaguely defined* [28]. Additionally, ethical, legal and societal concerns greatly affect the framework under which medical data may be used. The current US model encourages the use of de-identified, minimal risk medical data for research purposes, specifically data collected during routine treatment of patients. It is common for medical data collected in such a way to contain redundant, insignificant, incomplete or inconsistent data objects.

Rough Set Theory [29] is a mathematical tool for data analysis that has been used to address vagueness and inconsistencies present in datasets [30]. RST provides a systematic approach for analyzing data without implicit assumptions about relationships between covariates, an advantage that makes RST suitable for integration into medical applications [31]. RST operates on discretized numerical or symbolic data, and the information extracted from the dataset can be represented in the form of “if–then” decision rules—an intuitive representation that offers significant advantage over “black box” modeling approaches [32] and that increases accessibility and thus clinical credibility.

In the medical field, applications of RST focus mainly on the diagnosis and prognostication of diseases, where it has been demonstrated that RST is useful for extracting medical prognostic rules using minimum information. Tsumoto [33] argues that the concepts of approximation established in RST reflect the characteristics of medical reasoning, explaining why RST performs well in the medical field. For example, RST can be used to highlight non-essential prognostic factors in a particular diagnosis, thus helping to avoid redundant, superfluous or costly tests [34–38]. Recently, methods that combine survival analysis techniques and RST have been used to generate prognostic rules that estimate the survival time of a patient [24, 25].

3. Methods and Materials

3.1. Classical Rough Set Approach (CRSA)

Rough Set Theory, introduced by Pawlak in [29], provides methods for knowledge reduction by exploiting the relational indiscernibility of objects in an information system. Central to RST is the notion that an observed object has a certain amount of information associated with it. When considered in relation to a cohort of observed objects, this information is used to group similar objects into information granules. Together, the information provided by the set of observed objects can be generalized to describe the conditions required for membership in a concept class.

3.1.1. Notation

The methods of classical RST, hereafter referred to as the CRSA, act upon an information system of the form $S = (U, A, V, f)$, where U is a non-empty finite set of objects, called the universe. $A = C \cup \{d\}$ is a set of attributes that describe a given object in U, comprised of a set C of condition attributes and an optional decision attribute d. When d is present, the information system is a decision system, and is typically presented in table form. The set of all values, V, contains the value sets V_a, for every attribute $a \in A$. Given an object $x \in U$, $f\colon U \times A \rightarrow V$ maps the condition attribute of object x to its associated value $v = f(x, a) \in V_a$. A value attribute pair (a, v) for a given object is referred to as a descriptor.

In the CRSA, a data requirement is that the attribute values must be in discrete or categorical form. Table 1 provides an example of a discretized decision table, where six prognostic factors, as the condition attributes, describe seven patients. The decision attribute, presence of coronary disease in the patient, is represented by the binary attribute $d \rightarrow \{\text{Yes}, No\}$.

Once discretized, the objects in a decision table can be grouped according to their descriptors. For example, patients x_5 and x_6 have the same attribute values and are thus indiscernible from each other. In general, two objects $x_i, x_j \in U$ are indiscernible with respect to a set of condition attributes $B \subseteq C$ if $f(x_i, a) = f(x_j, a)\ \forall a \in B$. This relation is called an indiscernibility relation, given by $R(B) = \{(x_i, x_j) \in U\colon \forall a \in B, f(x_i, a) = f(x_j, a)\}$.

For example, the patients in Table 1 can be separated into four groups according to the indiscernibility relation $R(C)\colon X_1 = \{x_1\}, X_2 = \{x_2\}, X_3 = \{x_3, x_4, x_7\}, X_4 = \{x_5, x_6\}$. These groups of objects are referred to as equivalence classes, or conditional classes for $B \subseteq C$. An equivalence class for the decision attribute is called a decision class or concept, and in this example there are two groups: $Y_{No} = \{x_1, x_2, x_3\}$ and $Y_{Yes} = \{x_4, x_5, x_6, x_7\}$. The equivalence class specified by the object x_i with respect to $R(B)$ is denoted as $[x_i]_B$.

3.1.2. Set approximations

The goal of the CRSA is to provide a definition of a concept according to the attributes of the equivalence classes that contain objects that are known instantiations of the concept. As such, in a consistent decision table, membership in a conditional class implies membership in a particular decision class. In Table 1, $x \in X_4$ implies $x \in Y_{Yes}$. Membership in X_3, however, does not imply Y_{Yes} as $x_4, x_7 \in Y_{Yes}$ but $x_3 \in Y_{No}$. Thus Table 1 is inconsistent as $d(x_4, x_7) \neq d(x_3)$.

To represent an inconsistent decision table, the CRSA establishes an upper and lower approximation for each decision class, Y. The lower approximation is comprised of all objects that definitely belong to Y, while the upper approximation includes all objects that possibly belong to Y. It can be said that an object x_i definitely belongs to a concept Y if $[x_i]_C \subseteq Y$ and that x_i

possibly belongs to a concept Y if $[x_i]_C \cap Y \neq \emptyset$. Thus, the lower and upper approximations are defined as follows:

$$\underline{R}_B(Y) = \{x \in U \colon [x]_B \subseteq Y\} = \bigcup\{[x]_B \colon [x]_B \subseteq Y\}$$
$$\overline{R}_B(Y) = \{x \in U \colon [x]_B \cap Y \neq \emptyset\} = \bigcup\{[x]_B \colon [x]_B \cap Y \neq \emptyset\}$$
$$\overline{R}_B(Y) - \underline{R}_B(Y) = BND_B(Y)$$

The boundary region, $BND_B(Y)$, contains those objects that possibly, but not certainly, belong to Y. Conversely, the set $U - \overline{R}_B(Y)$ is the outside region containing those objects that certainly do not belong to Y. In our example, the lower and upper approximations for Y_{Yes} are $\underline{R}_C(Y_{Yes}) = X_4 = \{x_5, x_6\}$ and $\overline{R}_C(Y_{Yes}) = X_4 \cup X_3 = \{x_3, x_4, x_5, x_6, x_7\}$, and the boundary region contains the objects $BND_B(Y_{Yes}) = \{x_3, x_4, x_7\}$.

3.1.3. Reducts in the CRSA

Within a decision system, not all of the condition attributes may be required to define object-concept allocation. If, for an attribute subset $B \subseteq C$, the indiscernibility relation $R_B = R_C$, then the set approximations remain the same, the structure of the decision system is preserved and the attributes in $C - B$ are said to be dispensable. There may be many such subsets, but if B is minimal (does not contain any dispensable attributes), then B is termed a reduct. $\{Age,\ Smoker\}$ and $\{SystBP,\ HDL\}$ are two such reducts from our example decision table.

3.2. Dominance-Based Rough Set Approach (DRSA)

Under the DRSA [39] the relations between objects are no longer made by the indiscernibility relation as described in the CRSA [29]. Instead, the DRSA allows ordinal attributes with preference-ordered domains when a monotonic relationship exists between the attribute and the decision classes, for example when a “better” or “worse” value of an attribute leads to a “better” or “worse” decision class.

3.2.1. Overview of the DRSA

A decision table in the DRSA is expressed in the same way as the CRSA. To differentiate between attributes with and without preference order domains, those with a preference order are called criteria while those without are referred to as attributes, as in the CRSA.

In the DRSA the domain of criteria $a \in A$ is completely preordered by the outranking relation $\succeq_a$, representing the preference order of the domain. The outranking relation is also applicable for comparing two objects such that for $x_i, x_j \in U$, $x_i \succeq_a x_j$ means that x_i is at least as good as (outranks) x_j with respect to the criterion $a \in A$.

Commonly, the domain of a criteria a is a subset of real numbers, $V_a \subseteq R$ and the outranking relation is then a simple order “≥” on real numbers such that the following relation holds: $x_i \succeq_a x_j \Leftrightarrow f(x_i, a) \geq f(x_j, a)$. This relation is straightforward for gain-type criteria (the more, the better), and can be easily reversed for cost-type criteria (the less, the better).

Using Table 1 as an example, the decision class d is preference-ordered such that a positive diagnosis of coronary disease is assumed to be the “preferred” decision class. Criterion preference relations are then organized in the direction of the decision class; values which generally contribute to the incidence of coronary disease are preferred over those which indicate lower risk. For the criteria in Table 1, higher values are preferred to lower values—as in the case of

Appendix E (continued)

Age, *SystBP*, and *HDL*—and "Yes" is preferred to "No"—as in the case of *Smoker* and *Diabetic*. No such preference relation exists for *Gender*; as such, it is considered an attribute.

Let $T = \{1, \dots, n\}$ represent the domain V_d of the decision class d, by which the decision system is partitioned into n classes $Y = \{Y_t,\ t \in T\}$, where $Y_t = \{x \in U\colon f(x, d) = t\}$. Then, each object $x \in U$ is assigned to one and only one class Y_t. The decision classes are preference-ordered according to the decision maker, where the class indices represents the order of preferences, i.e. for all $r, s \in T$ such that for $r \geq s$ the objects from class Y_r are strictly preferred to the objects from class Y_s.

Upward and downward unions of classes are defined as:

$$Y_t^{\geq} = \bigcup_{s \geq t} Y_s \quad \text{and} \quad Y_t^{\leq} = \bigcup_{s \leq t} Y_s,\ \ s, t \in T$$

For any pair of objects $(x_i, x_j) \in U$, x_i dominates x_j with respect to a set of condition attributes $P \subseteq C$, denoted by $x_i\, D_P\, x_j$, if the following conditions are satisfied simultaneously:

$$\begin{aligned} &x_i \succeq_q x_j, \quad \text{for all critera } q \in P \\ &f(x_i, a) = f(x_j, a), \quad \text{for all attributes } a \in P \end{aligned}$$

The dominance relation defines two sets called dominance cones, where for each $x_i \in U$:

$$\begin{aligned} &D_P^+(x_i) = \{x_j \in U\colon x_j\, D_P\, x_i\}, \text{ representing the set of objects that dominates } x_i \\ &D_p^-(x_i) = \{x_j \in U\colon x_i\, D_P\, x_j\}, \text{ representing the set of objects dominated by } x_i \end{aligned}$$

Considering the dominance cones, the lower and upper approximations of the union of decision classes are defined as follows. The lower approximation $\underline{R}_P(Y_t^{\geq})$ represents objects that certainly belong to $Y_t^{\geq}$, such that there is no other object that dominates x and belongs to a decision class inferior to $Y_t^{\geq}$. Similarly, the lower approximation $\underline{R}_P(Y_t^{\leq})$ represents objects that certainly belong to $Y_t^{\leq}$, with no other object dominated by x and belonging to a decision class superior to $Y_t^{\leq}$. The upper approximations represent objects that possibly belong to one of the upward or downward unions of decision classes.

$$\begin{aligned} \underline{R}_P(Y_t^{\geq}) &= \left\{x \in U\colon D_P^+(x) \subseteq Y_t^{\geq}\right\} \\ \overline{R}_P(Y_t^{\geq}) &= \bigcup_{x \in Y_t^{\geq}} D_P^+(x) = \left\{x \in U\colon D_P^-(x) \cap Y_t^{\leq} \neq \emptyset\right\} \\ \underline{R}_P(Y_t^{\leq}) &= \left\{x \in U\colon D_P^-(x) \subseteq Y_t^{\leq}\right\} \\ \overline{R}_P(Y_t^{\leq}) &= \bigcup_{x \in Y_t^{\leq}} D_P^-(x) = \left\{x \in U\colon D_P^+(x) \cap Y_t^{\geq} \neq \emptyset\right\} \end{aligned}$$

Similar to the CRSA, the boundary regions are defined as:

$$\begin{aligned} BND_P Y_t^{\geq} &= \overline{R}_P(Y_t^{\geq}) - \underline{R}_P(Y_t^{\geq}) \\ BND_P Y_t^{\leq} &= \overline{R}_P(Y_t^{\leq}) - \underline{R}_P(Y_t^{\leq}) \end{aligned}$$

Using our example decision table, Table 1, and considering the full set of condition attributes, it can be seen that $x_4\, D_C\, x_3$, and furthermore $D_C^+(x_4) = \{x_3, x_4, x_7\}$, $D_C^-(x_4) = \{x_3, x_4, x_7\}$. Considering the dominance cones for all patients, the lower and upper approximations of the union of decision classes are $\underline{R}_C(Y_{Yes}^{\geq}) = \{x_5, x_6\}$, $\overline{R}_C(Y_{Yes}^{\geq}) = \{x_3, x_4, x_5, x_6, x_7\}$, $\underline{R}_C(Y_{No}^{\leq}) = \{x_1, x_2\}$, $\overline{R}_C(Y_{No}^{\leq}) = \{x_1, x_2, x_3, x_4, x_7\}$ and the boundary regions are $BND_C Y_{Yes}^{\geq} = BND_C Y_{No}^{\leq} = \{x_3, x_4, x_7\}$.

Appendix E (continued)

3.2.2. The variable consistency DRSA

The variable consistency DRSA (VC-DRSA) allows the decision maker to relax the strictness of the dominance relation, thus accepting a limited number of inconsistent objects in the lower approximation, according to a consistency level threshold, $l \in (0, 1]$. In practice, by selecting a consistency level l, a patient $x \in U$ becomes a member of a given decision class if at least $l{*}100\%$ of the patients dominating x also belong to that decision class. By allowing inconsistencies, the VC-DRSA avoids over fitting the training set and thus may be more effective in classifying new cases.

The lower approximations of the VC-DRSA-based model are represented as follows:

$$\underline{R}^l_P(Y_t^{\geq}) = \left\{x \in Y_t^{\geq} : \frac{|D_P^+(x) \cap Y_t^{\geq}|}{|D_P^+(x)|} \geq l\right\}$$

$$\underline{R}^l_P(Y_t^{\leq}) = \left\{x \in Y_t^{\leq} : \frac{|D_P^-(x) \cap Y_t^{\leq}|}{|D_P^-(x)|} \geq l\right\}$$

Continuing the example from Section 3.2.1, setting $l = 0.6$ moves the objects x_4 and x_7, previously included in the upper approximation $\overline{R}_C(Y^{\geq}_{Yes})$, to the lower approximation of class $Y^{\geq}_{Yes}$, i.e: $\underline{R}^{0.6}_C(Y^{\geq}_{Yes}) = \{x_4, x_5, x_6, x_7\}$. This follows from $\frac{|D_C^+(x_i) \cap Y_t^{\geq}|}{|D_C^+(x_i)|} = \frac{2}{3} \geq l$, for $i = 4, 5, 6, 7$.

3.2.3. Reducts in the VC-DRSA

For every subset of attributes $P \subseteq C$, the quality of approximation of the decision classes Y with respect to the attributes P, $\gamma_P(Y)$, is defined as the proportion among all objects in U of objects consistently defined with respect to the attributes P and the decision classes Y. Each subset $P \subseteq C$ such that $\gamma_P(Y) = \gamma_C(Y)$ is termed a reduct for both the VC-DRSA and DRSA.

$$\gamma_P(Y) = \frac{\left|U - \left\{\bigcup_{t \in T} BND_P Y_t^{\leq}\right\}\right|}{|U|} = \frac{\left|U - \left\{\bigcup_{t \in T} BND_P Y_t^{\geq}\right\}\right|}{|U|}$$

The subset of attributes $\{SystBP,\ HDL\}$ is an example of such a reduct since $\gamma_{SystBP,\,HDL}(Y) = \gamma_C(Y)$.

3.3. Decision rules

There are a number of methods available for induction of decision rules from the lower or upper approximations of the decision classes [40–42] or from reducts extracted from the decision table [43]. The rule induction methods used in this study are described in Section 3.5.1. Once decision rules have been induced, the collection of these rules can then be used to classify unseen objects—in the case of our example table, a new patient who may have cardiac disease.

In the CRSA, a decision rule has the form *if A then B*, or $A \rightarrow B$, where A is called the antecedent and B the consequent of the rule. The antecedent is a logical conjunction of descriptors and the consequent is the decision attribute or attributes suggested by the rule. For example, a CRSA decision rule induced from object x_1 from our example in Table 1 using the reduct $\{Age,\ Smoker\}$ would be: if *Age* = H and *Smoker* = No then *Coronary Disease* = No.

Formally, in the CRSA, a decision rule, generated from an object x' with respect to a set of condition attributes, $B \subseteq C$, is expressed as

$$\textit{if} \bigwedge_i \left(f(x, a_i) = x'_{a_i}\right) \textit{ then } x \in Y_{x'}$$

where $a_i \in B$ is an attribute found in the attribute set B, and $x'_{a_i} \in V_{a_i}$ and $Y_{x'}$ are the attribute values and decision class, respectively, of object x'.

In the DRSA, decision rules are induced from the lower approximations and the boundaries of the union of decision classes. From the lower approximations, two types of decision rules are considered. Decision rules generated from the P-lower approximation of the upward union of decision classes $Y_t^{\geq}$ are described by

$$\textit{if } \bigwedge_i \left(f(x, b_i) \geq r_{b_i}\right) \bigwedge \left(\bigwedge_j \left(f(x, a_j) = r_{a_j}\right) \right) \textit{ then } x \in Y_t^{\geq}$$

where $b_i \in P$ are criteria, $a_j \in P$ are attributes, $r_{b_i} \in V_{b_i}$ and $r_{a_j} \in V_{a_j}$. Decision rules generated from the P-lower approximation of the downward union of classes $Y_t^{\leq}$ are described by

$$\textit{if } \bigwedge_i \left(f(x, b_i) \leq r_{b_i}\right) \bigwedge \left(\bigwedge_j \left(f(x, a_j) = r_{a_j}\right) \right) \textit{ then } x \in Y_t^{\leq}$$

where $b_i \in P$ are criteria, $a_j \in P$ are attributes, $r_{b_i} \in V_{b_i}$ and $r_{a_j} \in V_{a_j}$. The boundaries $BND_P Y_t^{\geq}$ and $BND_P Y_t^{\leq}$ generate the following rules

$$\textit{if } \bigwedge_i \left(f(x, b_i) \geq r_{b_i}\right) \bigwedge \left(\bigwedge_j \left(f(x, b_j) \leq r_{b_j}\right) \right) \bigwedge \left(\bigwedge_k \left(f(x, a_k) = r_{a_k}\right) \right) \textit{ then } x \in Y_t^{\geq} \cup Y_t^{\leq}$$

where $b_i, b_j \in P$ are criteria, $a_k \in P$ are attributes, $r_{b_i} \in V_{b_i}$, $r_{b_j} \in V_{b_j}$ and $r_{a_k} \in V_{a_k}$ (note i and j are not necessarily different).

From our example in Table 1, the information from objects x_1 and $x_2 \in \underline{R}_C(Y_{No}^{\leq})$ yields the following rule: if *SystBP* ≤ H and *HDL* ≤ L then *Coronary Disease* = No.

3.4. Dataset description

3.4.1. SUPPORT dataset

The dataset used in this study is the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model dataset [44], a study of 9,105 terminally ill patients. SUPPORT enrolled patients, 18 years or older, who met specific criteria for one of nine serious illnesses, who survived more than 48 hours but were not discharged within 72 hours. Patients were followed such that survival and functional status were known for 180 days after entry. The result of the SUPPORT study is a prognostic model for 180-day survival estimation of seriously ill hospitalized adults based on cubic splines and a Cox regression model. Given the inclusion criteria (described in full in Appendix 1 of [12]), the dataset is ideal for the present research in regards to clinical applicability, completeness of data, and comparability of results.

We consider as condition attributes the variables used in the SUPPORT prognostic model equation [12] to ensure consistency. The SUPPORT variables include ten physiologic variables in addition to the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurologic function as recorded in the SUPPORT data. Attribute names, descriptions and value ranges are listed in Table 2.

The median survival time for the patients in the study is 223 days. Figure 1 shows the distribution of patients with respect to number of days until death. The SUPPORT study inclusion criteria was designed to include patients with 50% risk of death at 180 days; as seen in Table 2 death prior to 180s was observed in approximately 47% of patients.

General observations regarding the influence of condition attributes can be made by analyzing the distribution of time until death by attribute. For example, the box-whisker plot in Figure 2 shows that a significant portion (75%) of patients with coma or multi-organ system failure with malignancy (MOSF w/ Malig) do not survive longer than 180 days, but patients with congestive heart failure (CHF) or chronic obstructive pulmonary disease (CPD) tend to live longer than 180 days.

Note, also, that several *dzgroup* categories have a number of outliers, represented by circles in Figure 2. Whereas the information from these patients would be lost in a regression model, the RSA-based methods retain the information from these patients in the rule-generation and rule-application process. Given the number of outliers presented, however, it is reasonable to expect that a method that allows approximation (i.e. generalization) will be required to generate meaningful decision rules.

3.4.2. Data preprocessing

In its published form, the SUPPORT dataset contains 9,105 cases. Missing physiological attribute values are filled in with a standard fill-in value representing a normal physiological response, as provided by the SUPPORT authors in [44]. It is also worth noting that in the SUPPORT study, a patient for whom it was not possible to establish a Glasgow coma score was given a scoma value of zero. After missing data imputation, two cases are still incomplete; the remaining 9,103 cases were considered in the development of the prognostic models.

3.4.3. Discretization

Discretization is the process by which appropriate categorical ranges are found for variables with a continuous value range. This is a required step in the CRSA as the indiscernibility relations are computed on categorical condition attributes. In general, this step is not required for the DRSA, however in our study the chosen discretization method provides the necessary preference order relations for the DRSA and ensures directly comparable rule sets for all evaluated rule-based methods.

There are a number of methods available for unsupervised discretization that operate without input from the decision maker and are based only on the information available in the data table. In this work, however, discretization was primarily performed using the Acute Physiology and Chronic Health Evaluation (APACHE) III Scoring System [11], a clinically accepted scoring system designed to estimate the risk of death in ICU patients. In this sense, the use of the APACHE III scoring system represents a research-validated, clinically appropriate, expert discretization scheme. This choice is founded on the proposition that expert discretization via APACHE III will result in medically and contextually relevant classification rules and data collection requirements, thus increasing the clinical credibility of the proposed prognostic model.

In addition, APACHE III scores are designed to increase monotonically with respect to risk of death and thus provide the necessary preference relations for the DRSA. APACHE III scores for any given variable are close to zero for normal or only slightly abnormal values of that variable and increase according to increased severity of disease. For example, normal pulse rates of 50-99 bpm are given a score of 0, while elevated and lowered levels, 100-109 and 40-49 bpm respectively, are both given a score of 5. Thus, higher APACHE III scores are preferred to lower scores, as the higher scores indicate greater severity of disease and therefore greater risk of death within six months (considered the positive diagnosis).

For the rule-based methods considered in this study, the nine physiologic variables and the age variable were transformed to their representative APACHE III scores. The remaining physiologic variables not included in APACHE III—neurologic function, *scoma*, and blood gasses, *pafi*—were discretized using clinically accepted categorizations [45, 46]. The variable *hday* was discretized using the Boolean Reasoning Algorithm [47]. Table 3 shows the categories defined in this process. Higher values of each of these variables are preferred to lower values.

3.5. Algorithms and implementation details

3.5.1. Rough set rule induction and classification

Decision rules were obtained using the MODLEM [40, 41] and VC-DomLEM [42] algorithms for the induction of classical and dominance-based rough set rules. Both methods were applied to the discretized SUPPORT dataset. As both the MODLEM and VC-DomLEM algorithms generate a minimal set of decision rules using a minimal number of rule conditions, the inclusion of MODLEM allows for an evaluation of the impact of accounting for the preference order information.

Decision rules by sequential covering. The MODLEM and the VC-DomLEM algorithms utilize a heuristic strategy called *sequential covering* [48] to iteratively construct a minimal set of minimal decision rules. The sequential covering strategy successively constructs a set of decisions for each of the decision classes in a training set by selecting, at each iteration, the "best" decision rule, after which the training objects described by the rule conditions are removed. Subsequent iterations again select the best decision rule and remove the covered objects until reaching a stopping criteria or until all of the objects in the decision class are described by a rule in the rule set.

To ensure minimality, antecedent descriptors, called elementary conditions, of each rule are checked at each iteration and redundant elementary conditions are removed. Similarly, redundant rules are removed from the final rule set.

In both algorithms, decision rules are grown by consecutively adding the best available elementary condition to the rule. CRSA elementary conditions are evaluated in the MODLEM algorithm in terms of either the class entropy measure [49] or Laplacian accuracy [50]. Dominance-based elementary conditions are evaluated according to a rule consistency measure. VC-DomLEM provides three such measures; in this study the rule consistency, α, of a proposed rule, r_{Y_t}, suggesting assignment to decision class Y_t, is defined as

$$\alpha(r_{Y_t}) = \frac{\left|[\Phi(r_{Y_t})] \cap \underline{\mathrm{Y}}_t\right|}{\left|[\Phi(r_{Y_t})]\right|}.$$

Here $[\Phi(r_{Y_t})]$ indicates the set of objects described by the elementary conditions in r_{Y_t}. The elementary condition, *ec*, that is selected for inclusion is that which leads to the highest rule consistency measure $\alpha(r_{Y_t} \cup ec)$ when combined with the current set of elementary conditions in the proposed rule. In the event of a tie, the elementary condition providing greatest coverage of the new rule is selected, by $\left|[\Phi(r_{Y_t} \cup ec)] \cap \underline{\mathrm{Y}}_t\right|$. The rule consistency measure, α, can also be implemented in MODLEM to relax consistency requirements and allow more general rules to be induced. For further details on the MODLEM and VC-DomLEM algorithms, the reader is referred to [40–42].

MODLEM algorithm for CRSA decision rules. CRSA decision rules were obtained using the MODLEM algorithm as described in [40] and [41]. Decision rules were generated from the lower approximations with a rule consistency level $\alpha \geq m$. The rule syntax follows the presentation in Section 3.3.

VC-DomLEM algorithm for VC-DRSA decision rules. Dominance-based rules were obtained using the VC-DRSA as described in Section 3.2.2 and the VC-DomLEM algorithm as implemented in jMAF [51]. VC-DomLEM decision rules were generated from the lower approximation of each decision class. The syntax of the VC-DRSA decision rules is as shown in Section 3.3. Only decision rules with confidence greater than the consistency level, ie. decision rules with $\alpha \geq l$, are included in the classification model.

Parameter selection. In order to select the most appropriate models for comparison, the performance of the rough set based models was evaluated for varying levels of rule consistency, m and l, for the CRSA and VC-DRSA respectively. Classifier performance at a particular value of m or l is dataset-dependent; however, in general, values close to one provide rule sets that are more conservative in describing the training set objects, while values closer to zero provide rule sets that are more general. Thus, to find the appropriate balance between strict, descriptive models that are prone to overfitting and overly general models that provide little useful information, the RSA models were evaluated at $m, l = 0.1, 0.2, 0.4, 0.6, 0.8, 1.0$.

Classification. For the RSA-based models, a standard voting process [47] is used to allow all rules to participate in the decision process, arriving at a patient classification by majority vote. Each rule is characterized by two support metrics. The left hand side (LHS) support is the number of patients in the table whose attributes match the antecedent, while the right hand side (RHS) support indicates the number of patients matching the consequent of the rule. For a new, unseen patient, any rule whose antecedent descriptors match the patient descriptors "fires" by contributing as votes the RHS support for each decision class. Once all rules have "voted", the number of votes for each decision class is normalized against the total number of LHS support for all fired rules. The resultant ratio of RHS to LHS support is considered a frequency-based estimate of the probability that the patient belongs to the given decision class.

A final classification is therefore determined according to a threshold value, $\tau \in [0, 1]$. A patient is classified as not surviving six months if the estimated probability of death in six months is greater than τ. In the event of an estimated probability equal to τ, or in the absence of any fired rules (no rule matches the patient profile), classification is not possible and the patient is labeled *undefined*.

3.5.2. Comparative methods

To evaluate the performance of the RSA-based prognostic models, logistic regression, SVM, C4.5, and random forests were applied to the non-discretized SUPPORT dataset. Logistic regression was selected for its popularity in survival analysis and clinical settings [18, 52].

Support vector machines, originally presented in [53], find separating boundaries between decision classes after input vectors are non-linearly mapped into a high dimensional feature space. Support vector machines have been investigated in survival analysis applications [54] as they—similar to the RSA-based methods—automatically incorporate non-linearities and do not make *a priori* assumptions about factor interactions. SVM-based models are known to perform well at classification tasks, however they do not provide clinician-interpretable justification for their results [55]. Support vector machines were selected to evaluate whether the increased accessibility of the RSA-based methods involves a trade-off in accuracy.

C4.5 is a well known algorithm for generating a decision tree using information entropy to select the best splitting criteria at each node [56]. A decision tree built by C4.5 can be expressed as a set of if-then decision rules, thus providing a comparative decision rule based method. To ensure directly comparable rule sets, C4.5 was applied to the discretized SUPPORT dataset. However, as C4.5 provides methods for selecting appropriate cut-points in continuous attributes [57], a second model was also generated from the non-discretized data set.

Random forests is a popular ensemble classification method based on decision trees [58]. The random forests algorithm builds an ensemble of decision trees, where each tree is built on bootstrap samples of training data with a randomly selected subset of factors.

Each of these methodologies were applied to the non-discretized data set using the software package Weka 3.6.9 [59], with default parameters.

3.6. Performance evaluation methods

The performance of the models was tested by measuring the discriminatory power of both the *m*- and *l*-consistent decision rules sets when applied to the reserved testing data. For our notation, a classification of *d.6months = Yes* is referred to as a positive classification, and *d.6months = No* is negative. Sensitivity is defined as the fraction of patients who did not survive six months and are correctly classified by the model, or the fraction of true positive classifications of all test patients who did not survive six months. Conversely, specificity is defined as the fraction of patients who did survive six months and were correctly classified by the model, or the fraction of true negatives of all test patients who did survive six months.

The overall accuracy of the classification models is reported in terms of area under the Receiver Operating Characteristic (ROC) curve, or AUC (area under the curve). The ROC curve graphs the sensitivity of the classifier, or the true positive rate, versus $1 -$ specificity, the false positive rate, as the threshold probability, τ, for positive classification is varied from 0 to 1. The best overall classification performance is realized when AUC is equal to 1, while an AUC of 0.5 indicates a classifier performance no better than random selection. Best separation between decision classes is realized at the threshold corresponding to the point on the ROC curve closest to the point $(0, 1)$.

In order to select the most appropriate CRSA and VC-DRSA-based models for comparison, two performance issues related to the generated rule set were considered. The coverage of the classification model—i.e. the percentage of testing set patients for whom a classification is possible—for each *m* and *l* level was evaluated prior to selecting an appropriate level. To evaluate the number of rules that would fire for an unseen patient, we also collected information on the number of rules matching each test case patient for the evaluated levels of *m* and *l*.

Cohen's Kappa coefficient was computed for both the selected RSA-based models and the comparative models [60]. Cohen's Kappa coefficient is designed to measure the agreement between two classification methods, but it is commonly used to measure model performance by comparing a classifier with a random allocation of patients among the decision classes. A value of zero indicates classification accuracy equivalent to chance (zero disagreement).

To assess for significant differences in terms of AUC between the RSA-based methods and the aforementioned classification approaches, we applied the Wilcoxon Signed-Rank test [61]. The Wilcoxon Signed-Rank test is a non-parametric paired difference test that in this case is used to compare the performance of two classifiers by considering pairs of their AUC values over repeated runs.

4. Results

This section presents the results obtained using the CRSA, the VC-DRSA, logistic regression, SVM, C4.5 and random forests models for six-month life expectancy prognostication of terminally ill patients. The results are analyzed and compared.

To evaluate the performance of the prognostic models, a 5-fold cross validation procedure [62] was applied to repeatedly select training and testing sets. Cross validation is well known to provide a reasonable estimate of the generalization error of a prediction model. In 5-fold cross validation, the entire dataset is randomly divided into five subsets, or folds, and then each fold (20% of the dataset) is used once as a testing set, with the remaining folds (80%) used for training.

In order to select appropriate m and l values for the CRSA and VC-DRSA-based models, respectively, the performance of these models was evaluated first. AUC and coverage for each evaluated m and l level are shown in Table 4. Figures 3 and 4 display the number of rules that fire for each patient in the five testing folds for each m and l value. Based on these results, $m = l = 0.6$ was chosen as the rule consistency parameter for both CRSA and VC-DRSA-based classifiers for further evaluation with the comparative methods.

Table 5 describes the number of rules and the number of descriptors in each rule for the two rough set approach-based classifiers at the selected consistency level of 0.6. The average number of CRSA decision rules in the five rule sets generated by cross validation is 773 rules, with mean and maximum length of 3.65 and 8 descriptors, respectively. The VC-DRSA decision rules are on average slightly longer, with mean and maximum length of 6.85 and 13 elementary conditions, respectively. The mean total number of VC-DRSA rules is 1,095 rules.

The performance of all of the evaluated classification models is shown in Table 6, where Cohen's kappa coefficient [60] and AUC are reported for each classifier. Highest average kappa coefficient was achieved by VC-DRSA and logistic regression at $\bar{\kappa} = 0.35$. The random forest and C4.5 (using the pre-discretized SUPPORT data set) models obtained $\bar{\kappa} = 0.33$. The CRSA classifier and SVM approach achieved $\bar{\kappa} = 0.32$, followed by C4.5 (developed using the non-discretized SUPPORT data set) at $\bar{\kappa} = 0.31$.

The results of the Wilcoxon Signed-Rank test are presented in Table 7. Each table entry shows the p-value for the null hypothesis that there is no difference in the AUC between each pair of classifiers, when compared individually with the CRSA and VC-DRSA-based models.

5. Discussion

All of the methodologies show fair classification accuracy given that Kappa coefficients are in the range of 0.20 to 0.40 [63]. The results presented in Table 7 show no significant differences in the AUC when comparing the CRSA and VC-DRSA against the rest of the classification methods at a significance level of 0.05.

Clearly, m and l are critical values in determining model performance for both the CRSA and VC-DRSA. Together, Table 4 and Figures 3 and 4 demonstrate that selecting $m = l = 0.6$ balances the accuracy and coverage achieved by the rough set based classifiers against the amount of inconsistency allowed in each.

5.1. Interpretation and usability of decision rules

Clinical credibility in prognostic models depends in part on the ease with which physicians and patients can understand and interpret the results of the models, in addition to the accuracy of the information they provide. While the RSA-based prognostic models perform comparably to similar methods, by presenting the physician with a list of matched rules, the if-then decision rule approach offers significant advantages by increasing both the traceability of the model and the amount of information included in its results. This advantage is further increased in the case of the VC-DRSA, where dominance-based rules permit greater information density per rule by including attribute value ranges in each rule.

Table 8 contains the decision rules that fire for an example patient selected from the SUPPORT data set. This patient was 41 years old with a primary diagnosis of coma. The patient displayed moderate head injury on the Glasgow Coma Scale, elevated levels of creatinine (1.60 mg/dL) and respiratory rate (26 bpm), normal levels of sodium (133 mEq/L), low white blood cell count (1.90 cells/nL) and mean blood pressure of 107 bpm. Both the CRSA and VC-DRSA classifiers correctly predict that the patient will not survive six months (the patient in fact survived only 4 days).

The VC-DRSA classifier predicts *d.6months = Yes* with an associated score of 80%, based on the two rules (Rules 5 and 6). As can be seen in Table 8, Rule 5 isolates the combination of Coma and elevated creatinine and sodium levels as a key predictor of six-month survival. In the case of Rule 5, 51 patients in the training set have similar conditions as the example patient, of which 47 did not survive six months. On the other hand, Rule 6 somewhat counterbalances this prediction, pointing to 8 young patients with moderate coma who have been in the hospital less than 44 days, of whom all 8 survived six months.

The CRSA classifier provides a less specific prediction, classifying the example patient as not surviving six months with an associated score of 55%. Upon further investigation, the rules matching the example patient (Rules 1–4) are more general than the rules provided by the VC-DRSA classifier. Rules 1–3 provide general rules that point to the age, level of head trauma and primary diagnosis of the patient. Considering only these three rules, the associated score would be *d.6months = Yes* with a score of 54%, but this score is revised slightly by Rule 4 further in favor of *d.6months = Yes*. Rule 4 isolates normal average heart beat, high respiratory rate and low (and also very high) white blood cell counts.

For both the CRSA and VC-DRSA, a final prediction and associated score are presented by the classifier. This prediction is further supported by the set of rules from which said prediction derived. Thus, the gestalt survival expectation is presented without loss of contradictory information, providing the physician with both the prognostication as well as supporting and contradicting information. Furthermore, the rules clearly indicate the patient characteristics most relevant to their survival expectation. This increases the transparency and traceability of the classification process, strengthening the accessibility, and hence credibility, of the model.

The rules derived from the VC-DRSA, by including attribute value ranges for which the rule is valid, provide more information to the physician, further increasing the interpretability and utility of the life expectancy prediction. In a clinical setting, this set of rules serves to support clinical decisions for future treatment or palliative care strategies as well as to support the explanation of these decisions to the involved patient and their family.

This is in stark contrast to SVM, neural networks, and other black-box methods where very little insight is available to a decision maker as to how an outcome was predicted. While our results show similar performance in terms of accuracy between classification models, the RSA-based results are naturally expressed in terms of a set of decision rules, a benefit that is not present in logistic regression, random forests, or the mentioned black-box methods. As an ensemble method, the random forests method functionally reduces to a black-box style model, despite its use of decision trees.

Prognostic models based on C4.5 can be expressed in terms of the decision tree on which they are based or in the form of a set of decision rules. The benefits in terms of interpretability achieved by the decision rule format may be offset by the complexity of the tree growing and pruning methods used by C4.5 which limit the traceability of the model.

Additionally, rule-based prognostic models, including those based on the rough set approaches, are supported by a set of decision rules which do not individually involve all of the condition attributes. This offers the advantage of providing potentially acceptable results should a particular prognostic factor be difficult or too costly to ascertain for a patient [34].

5.2. Decision analysis for hospice referral

Consider the costs—economic, emotional and physical—associated with the decision to enter hospice care. These costs are justified for patients who either enter hospice care at the appropriate time or for those who do not enter hospice care when they could benefit from curative treatment. These cases represent true positive and true negative classifications. A higher emotional and physical cost is born by patients sent to hospice care but who ultimately survive six months—a false positive. The highest cost of all, emotionally, economically and physically is born by the patient and his or her family when costly treatment is prolonged for a patient who should have been referred to a hospice care program—a false negative. In this last case, some or all of the benefits of hospice care would be lost while the stresses and economic burden of aggressive treatment are endured.

In this light, the threshold parameter, τ (described in Section 3.5.1), can be seen as a representation of the patient and family's preference for hospice care treatment and their risk tolerance for a mistaken referral. The threshold parameter relates sensitivity to specificity and stipulates the required level of certainty for a positive classification. A higher threshold value requires a higher probability of not surviving six months for the classification of a patient as a hospice candidate, decreasing the sensitivity and increasing specificity (indicating a preference for continued treatment). Conversely, a lower threshold value increases sensitivity while reducing specificity, indicating a preference for avoiding the costly mistake of unnecessary treatment.

As this threshold value is a subjective matter and varies between physicians, patients and family members, one suggested approach [64] involves the measurement of the amount of regret the decision maker would have should an incorrect decision be made. As medical decisions must take into account the preferences of those ultimately affected by the decision, this application of regret theory allows for the formal treatment of those preferences by calculating the threshold value as a function of the measured anticipated regret.

6. Conclusions

This paper contributes to the growing body of research in RST—and its extensions—as a prognostic modeling framework and highlights the strengths of this approach in terms of accessibility. The CRSA and VC-DRSA are found to perform similarly to four common classification approaches, logistic regression, SVM, C4.5, and random forests, while also offering more information through a rule-based format. The intuitive structure of the rough set approaches, built on similarity relations and expressed in terms of if-then decision rules, offers both more insight into the model process and more opportunity for the knowledge extraction process to incorporate the personal preferences of those making and being affected by the decision.

The performance of the classifiers presented in this study, measured by AUC, is good but sub-optimal, indicative of a challenging problem in need of further research. The increased performance achieved by the variable consistency approach suggests a dataset of highly diverse patients. Future research will explore methods to improve the overall classifier performance and address this diversity by building localized models for patient subgroups using rough sets concepts to group patients with similar differentiating characteristics.

A recent study developed a six-month survival prognostic model primarily based on the Medicare Health Outcomes Survey responses of community-dwelling elderly patients [65]. This model, named the Patient-Reported Outcome Mortality Prediction Tool (PROMPT), achieved comparable AUC unsing only basic medical information, indicating that the performance of classification models for six-month survival is still a major issue for the targeted domain of hospice referral recommendation.

An important limitation of this study is that patient-specific disease progression over time is not considered, in part due to the static nature of the data set used. Future research must address the temporal aspect of disease progression, a consideration often missing in other prognostic models for hospice referral. The progression of a terminal illness is often highly non-linear by nature and generally does not present as a steady decline over time but rather as periods of relative stability marked by turning points of acute decline. A prognostic model that takes into account this temporal aspect may possibly provide both more accurate life expectancy prognoses and more useful information for palliative care planning.

Acknowledgments

This work was supported by the Department of Army under grant # W81 XWH-09-2-0175.

The authors thank Dr. Jurek Blaszczynski from the Laboratory of Intelligent Decision Support Systems at the Poznan University of Technology, Poznan, Poland, for his collaboration in providing valuable information about the Dominance Based Rough Set Approach.

Appendix E (continued)

7. References

[1] R. L. Kane, S. J. Klein, L. Bernstein, R. Rothenberg, J. Wales, Hospice Role in Alleviating the Emotional Stress of Terminal Patients and Their Families, Medical Care 23 (3) (1985) 189–197.

[2] W. Bulkin, H. Lukashok, Rx for Dying: the Case for Hospice, New England Journal of Medicine 318 (6) (1988) 376–378.

[3] N. J. Dawson, Need satisfaction in terminal care settings, Social science & medicine 32 (1) (1991) 83–87.

[4] N. A. Christakis, Timing of referral of terminally ill patients to an outpatient hospice., Journal of general internal medicine 9 (6) (1994) 314–20.

[5] J. M. Teno, J. E. Shu, D. Casarett, C. Spence, R. Rhodes, S. Connor, Timing of Referral to Hospice and Quality of Care: Length of Stay and Bereaved Family Members' Perceptions of the Timing of Hospice Referral, Journal of Pain and Symptom Management 34 (2) (2007) 120–125.

[6] N. A. Christakis, J. J. Escarce, Survival of Medicare Patients after Enrollment in Hospice Programs, New England Journal of Medicine 335 (3) (1996) 172–178.

[7] K. L. Lee, D. B. Pryor, F. E. Harrell, R. M. Califf, V. S. Behar, W. L. Floyd, J. J. Morris, R. A. Waugh, R. E. Whalen, R. A. Rosati, Predicting outcome in coronary disease statistical models versus expert clinicians, The American Journal of Medicine 80 (4) (1986) 553–560.

[8] J. C. Wyatt, D. G. Altman, Commentary: Prognostic models: clinically useful or quickly forgotten?, BMJ 311 (7019) (1995) 1539–1541.

[9] R. M. Dawes, D. Faust, P. E. Meehl, Clinical versus actuarial judgment, Science 243 (4899) (1989) 1668–1674.

[10] J. R. Beck, S. G. Pauker, J. E. Gottlieb, K. Klein, J. P. Kassirer, A convenient approximation of life expectancy (the DEALE): II. Use in medical decision-making, The American journal of medicine 73 (6) (1982) 889–897.

[11] W. A. Knaus, D. P. Wagner, E. A. Draper, J. E. Zimmerman, M. Bergner, P. G. Bastos, C. A. Sirio, D. J. Murphy, T. Lotring, A. Damiano, The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults, Chest 100 (6) (1991) 1619–1636.

[12] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, D. P. Wagner, The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults, Annals of Internal Medicine 122 (3) (1995) 191–203.

[13] N. A. Christakis, Predicting patient survival before and after hospice enrollment, Hospice Journal 13 (1998) 71–88.

[14] S. Gripp, S. Moeller, E. Bölke, G. Schmitt, C. Matuschek, S. Asgari, F. Asgharzadeh, S. Roth, W. Budach, M. Franz, et al., Survival prediction in terminally ill cancer patients by clinical estimates, laboratory tests, and self-rated anxiety and depression, Journal of Clinical Oncology 25 (22) (2007) 3313–3320.

[15] P. Glare, C. Sinclair, M. Downing, P. Stone, M. Maltoni, A. Vigano, Predicting survival in patients with advanced disease, European Journal of Cancer 44 (8) (2008) 1146–1156.

[16] I. Hyodo, T. Morita, I. Adachi, Y. Shima, A. Yoshizawa, K. Hiraga, Development of a Predicting Tool for Survival of Terminally Ill Cancer Patients, Japanese Journal of Clinical Oncology 40 (5) (2010) 442–448.

[17] P. K. J. Han, M. Lee, B. B. Reeve, A. B. Mariotto, Z. Wang, R. D. Hays, K. R. Yabroff, M. Topor, E. J. Feuer, Development of a Prognostic Model for Six-Month Mortality in Older Adults With Declining Health, Journal of Pain and Symptom Management 43 (3) (2012) 527–539.

[18] D. W. Hosmer, S. Lemeshow, N. Inc., Applied Survival Analysis: Regression Modeling of Time to Event Data, Wiley, New York, N.Y., 1999.

[19] R. Dybowski, V. Gant, P. Weller, R. Chang, Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm, The Lancet 347 (9009) (1996) 1146–1150.

[20] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, A. L. Dekker, Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy, Med Phys 37 (4) (2010) 1401–1407.

[21] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, W. T. Linde-Zwirble, Predicting hospital mortality for patients in the intensive care unit: A comparison of artificial neural networks with logistic regression models, Critical Care Medicine 29 (2).

[22] M. R. Segal, Features of Tree-Structured Survival Analysis, Epidemiology 8 (4) (1997) 344–346.

[23] S. S. Hwang, C. B. Scott, V. T. Chang, J. Cogswell, S. Srinivas, B. Kasimis, Prediction of Survival for Advanced Cancer Patients by Recursive Partitioning Analysis: Role of Karnofsky Performance Status, Quality of Life and Symptom Distress, Cancer Investigation 22 (5) (2004) 678–687.

[24] J. Bazan, A. Osmólski, A. Skowron, D. Ślçezak, M. Szczuka, J. Wroblewski, Rough set approach to the survival analysis, in: Rough Sets and Current Trends in Computing, Springer, 951, 2002.

[25] P. Pattaraintakorn, N. Cercone, K. Naruedomkul, Hybrid rough sets intelligent system architecture for survival analysis, in: W. M. Victor, O. Ewa, Owska, S. Roman, Owinski, Z. Wojciech (Eds.), Transactions on rough sets VII, Springer-Verlag, 206–224, 2007.

Appendix E (continued)

[26] D. Delen, G. Walker, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods., Artificial Intelligence in Medicine 34 (2) (2005) 113–27.

[27] K. J. Cios, G. W. Moore, Uniqueness of medical data mining, Artificial Intelligence In Medicine 26 (1-2).

[28] P. Simons, Critical Notice of Timothy Williamson, Vagueness, International Journal of Philosophical Studies 4 (1996) 321–327.

[29] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Norwell, MA, 1992.

[30] Z. Pawlak, Vagueness—A rough set view, Structures in Logic and Computer Science (1997) 106–117.

[31] P. Pattaraintakorn, N. Cercone, Integrating rough set theory and medical applications, Applied Mathematics Letters 21 (4) (2008) 400–403.

[32] A. Hart, J. Wyatt, Evaluating black-boxes as medical decision aids: issues arising from a study of neural networks, Informatics for Health and Social Care 15 (3) (1990) 229–236.

[33] S. Tsumoto, Modelling medical diagnostic rules based on rough sets, in: Rough Sets and Current Trends in Computing, Springer, 475–482, 1998.

[34] J. Komorowski, A. Øhrn, Modelling prognostic power of cardiac tests using rough sets, Artificial Intelligence in Medicine 15 (2) (1999) 167–191.

[35] P. Paszek, A. Wakulicz-Deja, Applying Rough Set Theory to Medical Diagnosing, Warsaw, Poland, 2007.

[36] M. Ningler, G. Stockmanns, G. Schneider, H.-D. Kochs, E. Kochs, Adapted variable precision rough set approach for EEG analysis., Artificial Intelligence in Medicine 47 (3) (2009) 239–61.

[37] H. Long-Jun, D. Li-pin, Z. Cai-Ying, Prognosis System for Lung Cancer Based on Rough Set Theory, in: Third International Conference on Information and Computing (ICIC), vol. 4, 7–10, 2010.

[38] C.-S. Son, Y.-N. Kim, H.-S. Kim, H.-S. Park, M.-S. Kim, Decision-making model for early diagnosis of congestive heart failure using rough set and decision tree approaches, Journal of Biomedical Informatics 5 (45) (2012) 999–1008.

[39] S. Greco, B. Matarazzo, R. Slowinski, Rough sets theory for multicriteria decision analysis, European Journal of Operational Research 129 (1) (2001) 1–47.

[40] J. Stefanowski, The rough set based rule induction technique for classification problems, in: Proc. 6th European Congress on Intelligent Techniques and Soft Computing, 7–10, 1998.

[41] J. Stefanowski, On combined classifiers, rule induction and rough sets, in: J. F. Peters (Ed.), Transactions on rough sets VI, Springer, 329–350, 2007.

[42] J. Blaszczyński, R. Słowiński, M. Szelag, Probabilistic Rough Set Approaches to Ordinal Classification with Monotonicity Constraints, in: E. Hllermeier, R. Kruse, F. Hoffmann (Eds.), Computational Intelligence for Knowledge-Based Systems Design, vol. 6178 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, ISBN 978-3-642-14048-8, 99–108, 2010.

[43] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, J. Wróblewski, Rough set algorithms in classification problem, in: Rough set methods and applications, Springer, 49–88, 2000.

[44] F. E. Harrell, SUPPORT Datasets, http://biostat.mc.vanderbilt.edu/wiki/Main/SupportDesc, 2010.

[45] S. C. Stein, Minor Head Injury: 13 Is an Unlucky Number, The Journal of Trauma and Acute Care Surgery 50 (4) (2001) 759–760.

[46] L. Martin, Reviews, Notes, and Listings: pulmonary Medicine: All You Really Need to Know to Interpret Arterial Blood Gases, Annals of Internal Medicine 8 (118) (1993) 656.

[47] J. G. Bazan, H. S. Nguyen, S. H. Nguyen, P. Synak, J. Wróblewski, Rough set algorithms in classification problem, in: Rough set methods and applications, Physica-Verlag GmbH, 49–88, 2000.

[48] J. Han, M. Kamber, J. Pei, Data Mining: Concepts and Techniques: Concepts and Techniques, The Morgan Kaufmann Series in Data Management Systems, Elsevier Science, 2011.

[49] J. W. Grzymala-Busse, J. Stefanowski, Three discretization methods for rule induction, International Journal of Intelligent Systems 16 (1) (2001) 29–38.

[50] P. Clark, R. Boswell, Rule induction with CN2: Some recent improvements (1991) 151–163.

[51] J. Błaszczyński, S. Greco, B. Matarazzo, R. Słowiński, M. Szelag, jMAF-Dominance-Based Rough Set Data Analysis Framework, Rough Sets and Intelligent Systems-Professor Zdzisław Pawlak in Memoriam (2009) 185–209.

[52] R. Bender, U. Grouven, Ordinal logistic regression in medical research, Journal of the Royal College of Physicians of London 31 (5) (1997) 546–551.

[53] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (3) (1995) 273–297.

[54] V. Van Belle, K. Pelckmans, S. Van Huffel, J. A. Suykens, Support vector methods for survival analysis: a comparison between ranking and regression approaches, Artificial Intelligence In Medicine 53 (2) (2011) 107–118.

[55] N. Barakat, A. P. Bradley, Rule extraction from support vector machines: A review, Neurocomputing 74 (1-3) (2010) 178–190.

[56] J. R. Quinlan, C4. 5: Programs for machine learning, vol. 1, Morgan kaufmann, 1993.

[57] J. R. Quinlan, Improved use of continuous attributes in C4.5, Journal of Artificial Intelligence Research 4 (1).

[58] L. Breiman, Random forests, in: Machine learning, 1, 5–32, 2001.

Appendix E (continued)

[59] C.-C. Chang, C.-J. Lin, LIBSVM: A library for support vector machines, ACM Trans. Intell. Syst. Technol. 2 (3) (2011) 27:1—-27:27.

[60] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement 20 (1) (1960) 37–46.

[61] F. Wilcoxon, Individual comparisons by ranking methods, Biometrics Bulletin 1 (6) (1945) 80–83.

[62] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection, 1995.

[63] A. Ben-David, Comparison of classification accuracy using Cohen's Weighted Kappa, Expert Systems with Applications 34 (2) (2008) 825–832.

[64] A. Tsalatsanis, I. Hoz, V. Andrew, B. Djulbegovic, I. Hozo, A. Vickers, B. Djulbegovic, A regret theory approach to decision curve analysis: a novel method for eliciting decision makers' preferences and decision-making., BMC Medical Informatics and Decision Making 10 (51) (2010) 51.

[65] P. K. J. Han, M. Lee, B. B. Reeve, A. B. Mariotto, Z. Wang, R. D. Hays, K. R. Yabroff, M. Topor, E. J. Feuer, Development of a Prognostic Model for Six-Month Mortality in Older Adults With Declining Health, Journal of Pain and Symptom Management 43 (3) (2011) 527–539.

Appendix E (continued)

8. Figures and Tables

Table 1: Example decision table

	Condition Attribute[a]						Decision Attribute
	c_1	c_2	c_3	c_4	c_5	c_6	d
Patient	Gender	Age	SystBP	HDL	Diabetic	Smoker	Coronary Disease
x_1	F	H	M	L	No	No	No
x_2	M	L	L	L	No	Yes	No
x_3	F	M	M	H	No	No	No
x_4	F	M	M	H	No	No	Yes
x_5	M	H	H	L	Yes	Yes	Yes
x_6	M	H	H	L	Yes	Yes	Yes
x_7	F	M	M	H	No	No	Yes

[a] Gender: Female/Male; Age: L = $[54, 59)$, M = $[59, 69)$, H = $[69, 74]$; SystBP: L =< 129, M = $[129 - 139]$, H = $(139 - 159]$; HDL: L =< 40 M = $[40 - 60]$, H => 60.

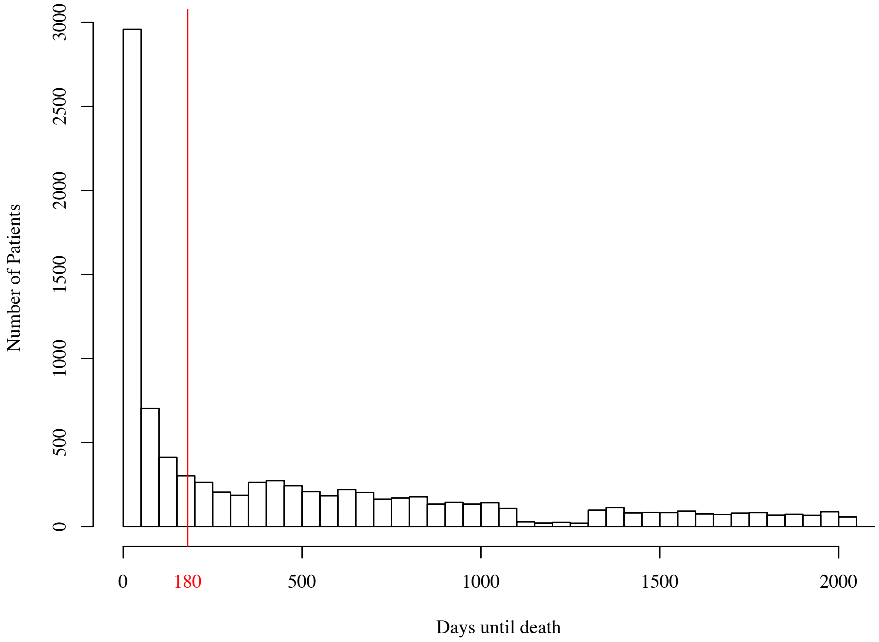
Appendix E (continued)

Table 2: Description of attributes from SUPPORT dataset

Variable Name	Description	Patient Distribution		
Numerical Condition Attributes		Range	Mean	Std. Dev
age	Age of the patient	18–101	62.65	15.59
alb	Serum albumin	0.4–29	2.95	0.87
bili	Bilirubin	0.1–63	2.55	5.32
crea	Serum creatinine	0.09–21.5	1.77	1.69
hday	Number of days in hospital at study entry	1–148	1.00	9.13
hrt	Heart Rate	0–300	97.16	31.56
meanbp	Mean arterial blood pressure	0–195	84.55	27.70
pafi	Blood gasses, $PaO_2/(.01 * FiO2)$	12–890.4	239.50	109.70
resp	Respiration rate	0–90	23.33	9.57
scoma	SUPPORT coma score, based on Glasgow coma scale	0–100	12.06	24.63
sod	Sodium	110–181	137.60	6.03
temp	Temperature in °C	31.7–41.7	37.10	1.25
wblc	White blood cell count	0–200	12.35	9.27
Categorical Condition Attributes		Patients	Percentage (%)	
dzgroup	Diagnosis Group:			
	ARF/MOSF w. Sepsis	3,515	38.60	
	CHF	1,387	15.23	
	Cirrhosis	508	5.56	
	Colon Cancer	512	5.62	
	Coma	596	6.54	
	COPD	967	10.60	
	Lung Cancer	908	9.97	
	MOSF w. Malignancy	712	7.81	
ca	Presence of cancer:			
	Yes	1,252	13.75	
	No	5,995	65.84	
	Metastasis	1,858	20.40	
Decision Attribute		Patients	Percentage (%)	
d.6months	Death occured within 6 months:			
	Yes	4,263	46.83	
	No	4,840	53.17	

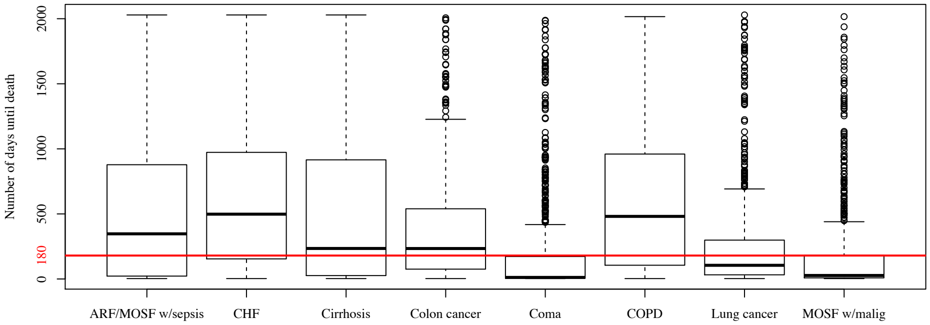
Appendix E (continued)

Figure 1: Distribution of patients with respect to number of days until death



Appendix E (continued)

Figure 2: Survival time in number of days vs. *dzgroup*



Appendix E (continued)

Table 3: Discretized attributes not in APACHE III

Attribute	Description	Categorization
scoma	Minor	$(*, 9]$
	Moderate	$(9, 44]$
	Severe	$(44, *)$
pafi	Normal	$[300, *)$
	Severe defect in gas exchange	$[200, 300)$
	Acute respiratory distress syndrome	$[0, 200)$
hday	Short	$(*, 44]$
	Long	$(44, *]$

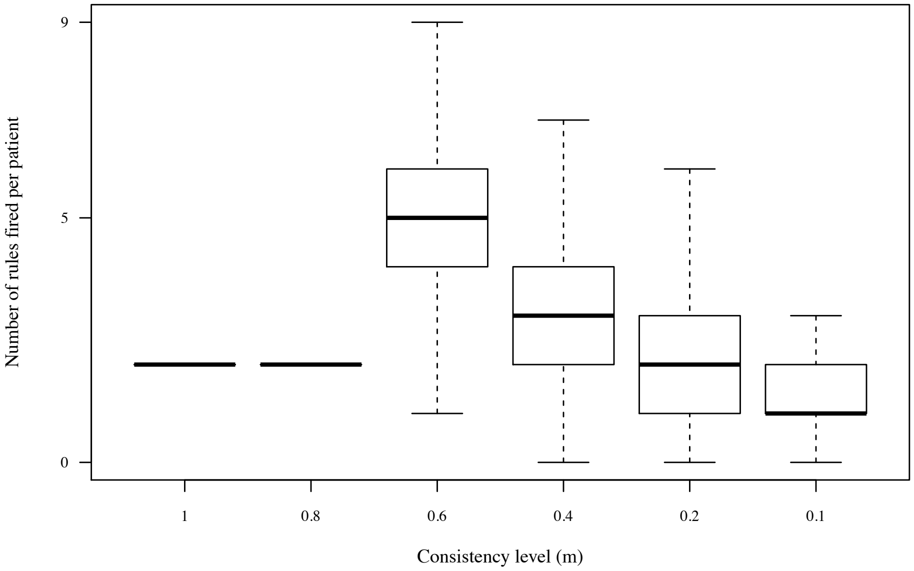
Appendix E (continued)

Table 4: AUC and coverage for MODLEM and VC-DomLEM algorithms with l and m-consistent rules

	CRSA		VC-DRSA	
m, l	AUC (%)	Coverage (%)	AUC (%)	Coverage (%)
0.1	66.46	100.00	72.80	99.88
0.2	66.46	100.00	72.79	99.87
0.4	68.88	100.00	72.77	99.65
0.6	69.74	97.41	71.73	98.72
0.8	64.19	86.72	70.93	76.85
1.0	61.58	80.08	65.59	35.89

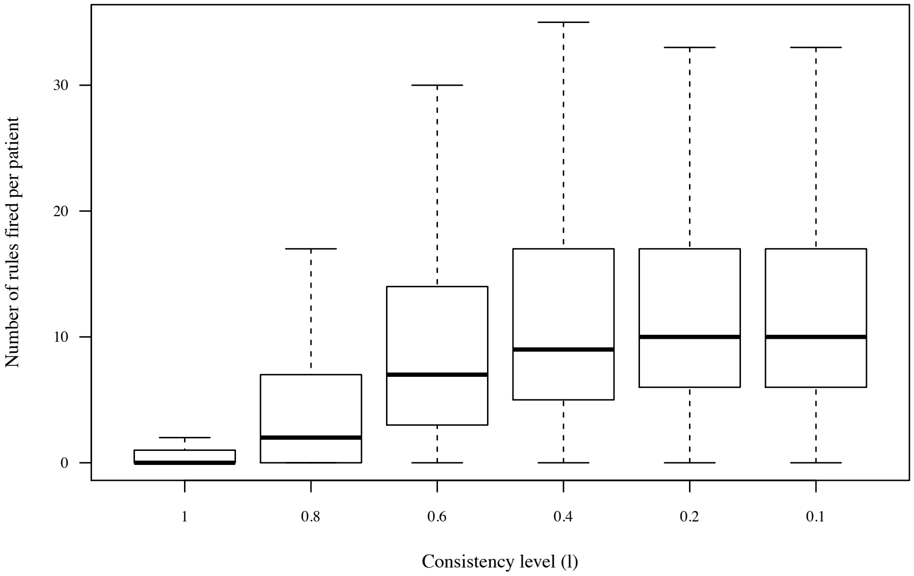
Appendix E (continued)

Figure 3: Number of rules fired in each test case for *m*-consistent MODLEM classifiers



Appendix E (continued)

Figure 4: Number of rules fired in each test case for *l*-consistent VC-DRSA classifiers



Appendix E (continued)

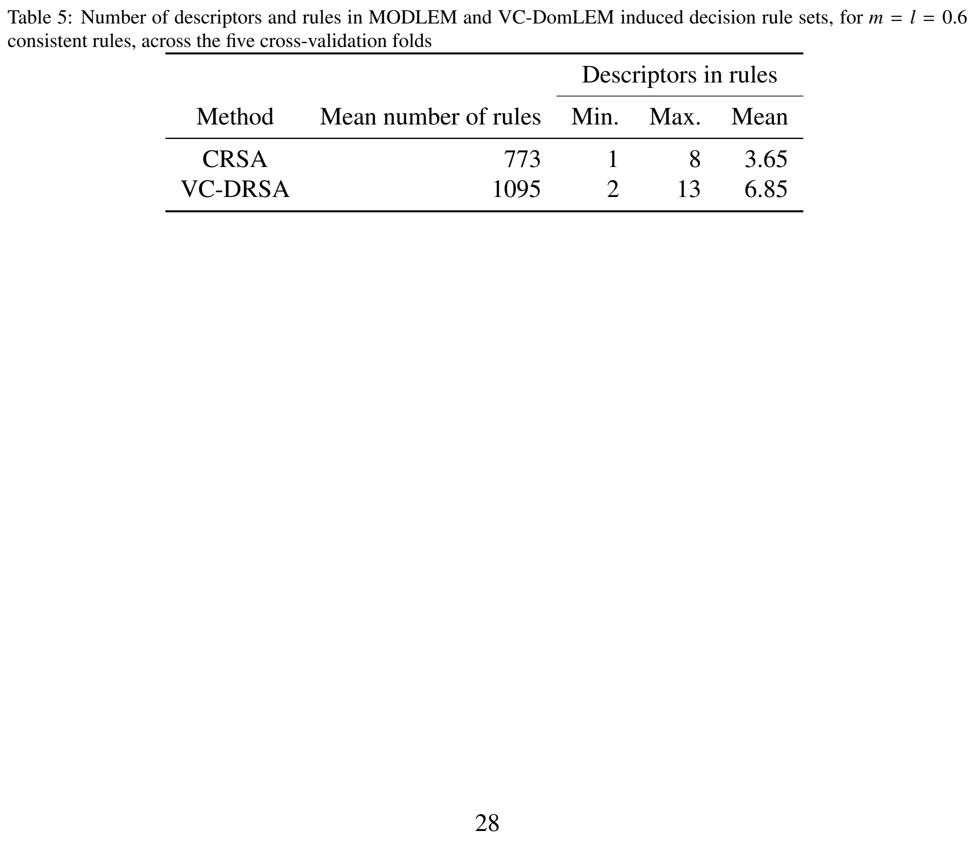


Table 5: Number of descriptors and rules in MODLEM and VC-DomLEM induced decision rule sets, for $m = l = 0.6$ consistent rules, across the five cross-validation folds

		Descriptors in rules		
Method	Mean number of rules	Min.	Max.	Mean
CRSA	773	1	8	3.65
VC-DRSA	1095	2	13	6.85

Appendix E (continued)

Table 6: Performance evaluation of the classification models: Logistic regression SVM C4.5 Random Forests CRSA and VC-DRSA

	Log. Reg.		SVM		C4.5		C4.5 *		Random Forest		CRSA**		VC-DRSA***	
Testing fold	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)	Kappa	AUC (%)
1	0.36	75.00	0.34	73.80	0.35	67.30	0.33	68.50	0.34	74.30	0.35	69.95	0.35	71.54
2	0.36	73.70	0.31	72.80	0.25	64.20	0.32	68.10	0.35	72.50	0.33	70.16	0.38	72.53
3	0.36	73.30	0.32	72.10	0.35	70.00	0.35	70.00	0.31	71.10	0.27	67.58	0.32	70.48
4	0.36	75.30	0.33	74.50	0.33	67.30	0.35	69.60	0.35	73.80	0.34	71.78	0.37	73.64
5	0.33	73.70	0.32	72.10	0.27	64.70	0.32	66.30	0.31	71.10	0.31	69.21	0.32	70.44
Mean	0.35	74.20	0.32	73.06	0.31	66.70	0.33	68.50	0.33	72.56	0.32	69.74	0.35	71.73
Std. Dev.	0.01	0.89	0.01	1.06	0.05	2.34	0.02	1.45	0.02	1.49	0.03	1.53	0.03	1.37

[*] C.45 with APACHE III discretized scores

[**] CRSA with MODLEM algorithm (m = 0.6)

[***] VC-DRSA with VC-DomLEM algorithm (l = 0.6)

Appendix E (continued)

Table 7: Wilcoxon signed-rank test and *p*-values for comparison of CRSAs and VC-DRSA with other classifiers

	Log. Reg.	SVM	C4.5	C4.5*	Random Forest	CRSA**	VC-DRSA***
CRSA**	0.06	0.06	0.06	0.06	0.06	—	0.06
VC-DRSA***	0.06	0.06	0.06	0.06	0.125	0.06	—

* C.45 with APACHE III discretized scores

** CRSA with MODLEM algorithm (*m* = 0.6)

*** VC-DRSA with VC-DomLEM algorithm (*l* = 0.6)

Appendix E (continued)

Table 8: Selected decision rules from the CRSA using MODLEM and the VC-DRSA using VC-DomLEM

			RHS Support	
	CRSA Rules using MODLEM	LHS	*d.6months = No*	*d.6months = Yes*
1.	If *age_score*[a] = 0	969	593 (61%)	376 (39%)
2.	If *scoma* = Moderate	1016	399 (39%)	617 (61%)
3.	If *dzgroup* = Coma	465	119 (26%)	346 (74%)
4.	If *hrt_score*[b] = 0 AND *resp_score*[c] = 6 AND *wbc_score*[d] = 5	47	11 (23%)	36 (77%)
	VC-DRSA Rules using VC-DomLEM			
5.	If *dzgroup* = Coma AND *crea_score*[e] ≥ 4 AND *sod_score*[f] ≥ 2	51	4 (8%)	47 (92%)
6.	If *dzgroup* = Coma AND *scoma* ≤ Moderate AND *hday* ≤ Short AND *age_ score*[a] ≤ 0	8	8 (100%)	0 (0%)

[a] *age_score:* 0 = (*age* ≤ 44)

[b] *hrt_score:* 0 = (50 ≤ *hrt* ≤ 99)

[c] *resp_score:* 6 = (25 ≤ *resp* ≤ 34)

[d] *wbc_score:* 5 = ((1 ≤ *wbc* ≤ 2.9) or (*wbc* ≥ 25))

[e] *crea_score:* ≥ 4 = (*crea* ≥ 1.5)

[f] *sod_score:* ≥ 2 = ((*sod* ≤ 134) or (*sod* ≥ 155))

Appendix F: Towards a Patient-centered Classification Model for Hospice Referral

Appendix F (continued)

Towards a patient-centered classification model for hospice referral

Eleazar Gil-Herrera

Department of Industrial and Management Systems Engineering, University of South Florida, Tampa, FL 33620, USA

Abstract

We introduce a methodology for developing a patient-centered classification model to determine potential hospice candidates in a population of terminally ill patients. In a patient-centered approach, those patients whose characteristics differ from the rest of the population may require different models to determine their classification. This is in contrast to population-based models that induce a single model to be applied for all patients.

In the data analysis phase of the proposed methodology we use the Object Relate Reducts (ORR) to identify indispensable patient characteristics that differentiate it from other patients having a different outcome. Since we consider condition attributes with preference-ordered domains, the ORRs are obtained using the Dominance Based Rough Set Approach (DRSA). These type of reducts are called Dominance Based Object Related Reducts (DORR).

The DORRs are used to construct subgroups of patients with similar characteristics in terms of the condition attributes necessary for classification. The collection of decision rules relative to each subgroup is used to classify new patients. The performance of the proposed methodology is compared with commonly known rough set-based methodologies such as the MODLEM and VC-DOMLEM algorithms.

1. Introduction

1.1. Characteristics of prognostic models for life expectancy

Life expectancy prognostication is particularly valuable for terminally ill patients since an accurate prognostication enables them to initiate end-of-life discussions and choose the most desired management strategy for the remainder of their lives.

Prognostic models for estimating life expectancy are developed to improve physicians' survival estimations. These models are the result of a feature selection process and include variables with high predictor-response correlation. Each variable in the model has to show independent

October 29, 2013

statistical significance. This variable-centered methodology represents the prognostic model in terms of an equation with coefficients associated with each variable.

Well-known mortality prognostic models [4, 5], show good prediction accuracy when they are applied to a population of critically ill patients. However, research shows limitations on their ability to predict outcomes when applied to individual patients [5]. A recent study [7] shows that it is not sufficient for a predictor variable to have statistical significance in a global model to be considered useful for individual patient prognosis. Instead, [7] states that each variable should be evaluated in terms of its role in identifying patients with differential response to a given treatment. Therefore, it is important to identify variables that differentiate groups of patients and at the same time are relevant in making decisions that better benefit an individual patient.

1.2. Patient-centered analysis in clinical data

Patient-centered analysis identify the most relevant factors that drive clinical decisions for an individual patient, in contrast to the commonly used population-wide models that are constructed to perform well on average on all future cases. Several studies [2, 7–10] have shown that patient-centered methodologies improve the accuracy of the prognostic model and assist in identifying profiles of patients with high risk of mortality.

The new vision of personalized health care [11] requires new methods for developing patient-centered prognostic models resulting in the selection of specific and appropriate treatment for each patient. In addition, to meet the needs of physicians and patients, prognostic models must have clinical credibility [12, 16]. That is, in addition to accurate prognostication, a model should be traceable in its structure, allowing complete insight to the prognostic process and its results should be interpretable, thus facilitating explanation of the prognosis.

In this paper we present the design and development of a new methodology based on Rough Set Theory (RST) [13], for analyzing clinical datasets and develop a patient-centered classification model. Since we consider condition attributes with preference-ordered domains, we use the Dominance based rough set approach (DRSA) [15] to obtain the Object Related Reducts (ORR) [14] for our dataset. We call this type of reducts as the *Dominance Based Object Related Reducts* (DORR).

Central to this methodology is the identification of indispensable patient characteristics that differentiate it from other patients having a different outcome. We separate the population of patients into smaller subgroups based on those indispensable characteristics. To classify new patients our methodology considers decision rules pertaining to their corresponding subgroup only. The performance of the proposed classification model is compared with other RST-based methods discussed in [16].

The rest of the paper is organized as follows: Section 2, methods and materials, describes the theoretical basis of the proposed methodology and the dataset used to demonstrate our method. Section 3 presents the results obtained as well as the comparison results with the selected methodologies. Finally in Section 4 we present the discussion of results and conclusions of this work.

2. Methods and materials

Our methodology is based on Rough Set Theory (RST) [13], a mathematical tool designed to analyze datasets. The basic concepts of RST consider the relation of objects in a dataset to group similar objects into granules of information called equivalence classes. RST-based tools in data analysis rely on their advantages to analyze datasets without previous assumptions and provide readily interpretable results in the form of decision rules.

2.1. Basic notation of Rough Set Theory

RST represents a dataset as an information system defined by $S = (U, A, V, f)$ where U, called the universe, is a non-empty finite set of objects that represents real life entities. The set A represents a non-empty finite set of attributes called the condition attributes. For every attribute $a \in A$, the function $f : U \rightarrow V_a$ makes a correspondence between an object $u \in U$ to an attribute value V_a called the value set of a. For datasets that include an outcome variable, RST defines a decision system as $DS = (U, A \bigcup d, V, f)$, where $d \notin A$ is called the decision attribute which represents the outcome variable. The domain of the decision attribute defines equivalence classes called decision classes. For binary decision attributes, two decision classes are defined and are represented as Y_0 and Y_1.

2.2. Discernibility relation

Data analysis in RST is based on relations between objects in a dataset. These relations considers similarities or differences between objects. In RST, similarities are represented by a discernibility relation, mathematically defined as:

$$DIS_{DS}(B) = \left\{(u, u') \in U^2, \exists a \in B : f(u, a) \neq f(u', a) \text{ and } f(u, d) \neq f(u', d)\right\} \forall B \in A$$

A discernibility matrix Dm_A is constructed by exploring the differences between objects in a decision system. Each cell of the matrix, $Dm_A(u, u')$, contains the set of attributes whose values differentiate a pair of objects $u, u' \in U$, i.e:

$$Dm_A(u, u') = \{a \in B : f(u, a) \neq f(u', a) \text{ and } f(u, d) \neq f(u', d)\}$$

Table 1 provides an example of a decision system DS with four prognostic factors (a_1, a_2, a_3 and a_4), as the condition attributes to describe five male patients (u_1, u_2, u_3, u_4 and u_5). The decision attribute represents the outcome of a fertility test performed on each patient, and is represented by the binary attribute $d \rightarrow \{Normal, Altered\}$.

Table 1: Example of a decision system

	Condition Attribute[a]				Decision Attribute
	a_1	a_2	a_3	a_4	d
Patient	Age	ChildDisease	Smoking	HoursSitting	FertilityTest
u_1	L	1	1	0	Normal
u_2	L	0	0	2	Normal
u_3	L	0	0	0	Altered
u_4	L	1	1	2	Altered
u_5	M	1	3	3	Altered

[a] Age: L =< 40, M = $(40, 60]$; ChildDisease (chicken pox, measles, mumps, polio) 0 = *No*, 1 = *Yes*; Smoking: 0 = Never, 1 = Occasionally, 2 = Frequently, 3 = Daily; HrSitting: 0 = $[0, 3]$, 1 = $(3, 6]$, 2 = $(6, 8]$, 3 = > 8

Using this information, we can generate the discernibility matrix Dm_A shown in Table 2. For example, for patients $u_1 \in Y_{Normal}$ and $u_3 \in Y_{Altered}$, the cell $Dm_A(u_1, u_3) = \{a_2, a_3\}$, since $(f(u_1, a_2) = 1) \neq (f(u_3, a_2) = 0)$ and $(f(u_1, a_3) = 1) \neq (f(u_3, a_3) = 0)$. Therefore, $\{a_2, a_3\}$ are the attributes that differentiate u_1 and u_3.

Table 2: Example of discernibility matrix

	u_1	u_2	u_3	u_4	u_5
u_1			a_2, a_3	a_4	a_1, a_3, a_4
u_2			a_4	a_2, a_3	A
u_3	a_2, a_3	a_4			
u_4	a_4	a_2, a_3			
u_5	a_1, a_3, a_4	A			

2.2.1. Object related reducts

One can consider the information form the discernibility matrix to obtain a minimum set of attributes (reduct) that distinguish a particular object $u \in U$ from the rest of the objects that belong to a different decision class. This type of reduct is called an object related reduct (ORR) and is defined as:

$$\forall u_i \in U : f(u_i, d) \neq f(d, u_j) \implies \exists a_k \in ORR_u : f(u_i, a_k) \neq f(u_j, a_k) \text{ where } u_i \neq u_j$$

To obtain the ORR's for each object, we apply the *discernibility function* [17] relative to each row of the discernibility matrix. A *discernibility function* relative to a row *i*, is a boolean function of the *m* condition attributes that appear in row *i* of the discernibility matrix Dm_A, i.e.:

$$f_i(a_1\text{'}, \ldots, a_m\text{'}) = \bigwedge \left(\bigvee D_m\text{'}(u_i, u_j) | j \leq |U|, D_m\text{'}(u_i, u_j) \neq \emptyset \right),$$

where, $D_m\text{'}(u_i, u_j) = \{a\text{'} | a \in D_m(u_i, u_j)\}$

The resultant prime implicants from each discernibility function f_i are the ORR's corresponding to each object $u_i \in U$. In our example we have:

- $u_1 : f_1(a_1, a_2, a_3, a_4) = (a_2 \vee a_3) \wedge (a_4) \wedge (a_1 \vee a_3 \vee a_4) \equiv (a_2 \wedge a_4) \vee (a_3 \wedge a_4)$, then $ORR_{u_1} = \{a_2, a_4\}, \{a_3, a_4\}$

- $u_2 : f_2(a_1, a_2, a_3, a_4) = (a_4) \wedge (a_2 \vee a_3) \wedge (a_1 \vee a_2 \vee a_3 \vee a_4) \equiv (a_2 \wedge a_4) \vee (a_3 \wedge a_4)$, then $ORR_{u_2} = \{a_2, a_4\}, \{a_3, a_4\}$

- $u_3 : f_3(a_2, a_3, a_4) = (a_2 \vee a_3) \wedge (a_4) \equiv (a_2 \wedge a_4) \vee (a_3 \wedge a_4)$, then $ORR_{u_3} = \{a_2, a_4\}, \{a_3, a_4\}$

- $u_4 : f_4(a_2, a_3, a_4) = (a_4) \wedge (a_2 \vee a_3) \equiv (a_2 \wedge a_4) \vee (a_3 \wedge a_4)$, then $ORR_{u_4} = \{a_2, a_4\}, \{a_3, a_4\}$

- $u_5 : f_5(a_1, a_2, a_3, a_4) = (a_1 \vee a_3 \vee a_4) \wedge (a_1 \vee a_2 \vee a_3 \vee a_4) \equiv (a_1) \vee (a_3) \vee (a_4)$, then $ORR_{u_5} = \{a_1\}, \{a_3\}, \{a_4\}$

For each patient, the ORRs account for the minimal set of condition attributes that preserve the differences of that patient with respect to the others in a different decision class.

2.2.2. Grouping objects based on Object Related Reducts

Using the ORR's, we construct subgroups of objects by considering the ones having the same ORR, where, each subgroup is described by a unique set of attributes. The ORRs also guarantee that the subgroups are constructed avoiding the use of redundant attributes.

Let $RED(U)$ be the set of all ORR's obtained from a decision system. In our example, $RED(U) = \{\{a_1\}, \{a_3\}, \{a_4\}, \{a_2, a_4\}, \{a_3, a_4\}\}$. Then, for each element $S \in RED(U)$, we can construct the following subgroups:

$$S_{a_1} = S_{a_3} = S_{a_4} = \{u_5\}$$

$$S_{a_2,a_4} = S_{a_3,a_4} = \{u_1, u_2, u_3, u_4\}$$

Using the information of the subgroups, one can observe that for patient u_5, the only required information to distinguish him from patients with *Normal* fertility test results is his Age (a_1). Alternatively we can use information about his smoking habits (a_3) or the number of hours he spends sitting per a day (a_4). On the other hand, for patients u_1, u_2, u_3 and u_4, besides the information on the number of hours they spend sitting (a_4), information about their smoking habits (a_3) or the occurrence of a disease in their childhood (a_2) is required.

2.2.3. Decision rules based on ORR

Decision rules are generated for each subgroup S_B. As an example, considering the subgroup $S_{a_4} = \{u_5\}$ and $S_{a_3,a_4} = \{u_1, u_2, u_3, u_4\}$, then, the following rules are obtained from Table 1. The objects that support the decision rule appear in parenthesis :

For S_{a_4}:

if $f(u, HoursSitting) = 3$ then $u \in Y_{Altered}$ (u_5)

Appendix F (continued)

For S_{a_3,a_4}:

$$\text{if } f(u, Smoking) = 1 \text{ and } f(u, HoursSitting) = 0 \text{ then } x \in Y_{Normal} \qquad (u_1)$$

$$\text{if } f(u, Smoking) = 0 \text{ and } f(u, HoursSitting) = 2 \text{ then } x \in Y_{Normal} \qquad (u_2)$$

$$\text{if } f(u, Smoking) = 0 \text{ and } f(u, HoursSitting) = 0 \text{ then } x \in Y_{Altered} \qquad (u_3)$$

$$\text{if } f(u, Smoking) = 1 \text{ and } f(u, HoursSitting) = 2 \text{ then } x \in Y_{Altered} \qquad (u_4)$$

2.2.4. Dominance based object related reducts (DORR)

The classical RST does not consider information about preference orders for classification. However, this information is particularly valuable in many practical problems that involve the evaluation of objects based on preference ordered domains. Blaszczynski et al. [15] present a new approach called the Dominance Based Rough Set Approach (DBRA) that consider attributes with preference-ordered domains (criteria) in both the condition and decision attributes.

When the domain of a criteria a is a subset of real numbers $V_a \subseteq R$, the outranking relation is then a simple order ”≥” on real numbers such that the following relation holds: $u_i \geq_a u_j \iff f(u_i, a) \geq f(u_j, a)$. This relation is straightforward for gain-type criteria (the more, the better), and can be easily reversed for cost-type criteria (the less, the better).

Using Table 1 as an example, the decision class d is preference-ordered such that an altered result in the fertility test is assumed to be the preferred decision class. The attribute-preference relations are then organized in the direction of the decision class; values which generally contribute to the abnormality in the test are preferred over those which indicate normality. For the criteria in Table 1, higher values are preferred to lower values.

Considering the dominance principle, we redefine the discernibility matrix Dm_A as follows, $\forall(u, u^{'} \in U)$:

Appendix F (continued)

$$Dm_A = \begin{cases} a \in A : f(u,a) > f(u\text{'},a) \text{ and } f(u,d) > f(u\text{'},d) & \text{if } a \text{ is criterion} \\ a \in A : f(u,a) < f(u\text{'},a) \text{ and } f(u,d) < f(u\text{'},d) & \text{if } a \text{ is criterion} \\ a \in A : f(u,a) \neq f(u\text{'},a) \text{ and } f(u,d) \neq f(u\text{'},d) & \text{if } a \text{ is attribute} \end{cases}$$

Table 3 presents the discernibility matrix considering the preference order of the attribute domains.

Table 3: Example of discernibility matrix considering criteria

	u_1	u_2	u_3	u_4	u_5
u_1				a_4	a_1, a_3, a_4
u_2				a_2, a_3	A
u_3					
u_4	a_4	a_2, a_3			
u_5	a_1, a_3, a_4	A			

Note in Table 3, the row and column corresponding to patient u_3 are now empty, compared to the discernibility matrix presented in Table 2. Attributes a_2, a_3 and a_4 are now removed due to information that is inconsistent with the dominance principle. That is, $(f(u_3, ChildDisease) = 0) < (f(u_1, ChildDisease) = 1)$ and $u_1 \in Y_{Normal}$. The same situation is observed for $(f(u_3, HoursSitting) = 0) < (f(u_2, HoursSitting) = 2)$ as $u_2 \in Y_{Normal}$.

The Dominance-based object related reducts (DORR) are obtained as follows:

- $u_1 : f_1(a_1, a_2, a_3, a_4) = (a_4) \wedge (a_1 \vee a_3 \vee a_4) \equiv (a_4)$, then $DORR_{u_1} = \{a_4\}$
- $u_2 : f_2(a_1, a_2, a_3, a_4) = (a_2 \vee a_3) \wedge (a_1 \vee a_2 \vee a_3 \vee a_3) \equiv (a_2 \wedge a_3)$, then $DORR_{u_2} = \{a_2, a_3\}$
- $u_3 : \emptyset$, then $DORR_{u_3} = \emptyset$
- $u_4 : f_3(a_2, a_3, a_4) = (a_4) \wedge (a_2 \vee a_3) \equiv (a_2 \wedge a_4) \vee (a_3 \wedge a_4)$, then $DORR_{u_4} = \{a_2, a_4\}, \{a_3, a_4\}$
- $u_5 : f_4(a_1, a_2, a_3, a_4) = (a_1 \vee a_3 \vee a_4) \wedge (a_1 \vee a_2 \vee a_3 \vee a_4) \equiv (a_1) \vee (a_3) \vee (a_4)$, then $DORR_{u_5} = \{a_1\}, \{a_3\}, \{a_4\}$

2.2.5. Grouping objects based on Dominance Object Related Reducts

DORR's subgroups are obtained using the same grouping process described for the ORR's in section 2.3:

$$S_{a_1} = \{u_5\}, S_{a_3} = \{u_5\}, S_{a_4} = \{u_1, u_5\}$$

$$S_{a_2,a_3} = \{u_2\}, S_{a_2,a_4} = S_{a_3,a_4} = \{u_4\}$$

The subgroups obtained from the DORRs have the following property:

Property. Let $B, B' \subseteq A$ and $B, B' \in RED(U)$. If $B \subset B^{'}$, then the subgroups S_B and $S_{B'}$ are disjoint, i.e. $S_B \cap S_{B'} = \emptyset$.

Proof. Assuming $B \subset B^{'}$ then, $B \cap B^{'} = B$. We need to proof that $S_B \cap S_{B'} = \emptyset$.

Suppose $S_B \cap S_{B'} \neq \emptyset$, then, $\exists u \in S_B$ and $u \in S_{B'}$. This implies that $B \in DORR_u$ and $B^{'} \in DORR_u$. However, the $DORR_u$ is the prime implicant of the boolean function that includes the term $B \wedge B^{'}$, which implies that $B \not\subset B^{'}$, contradicting our assumption.

In our example, $B = \{a_4\}$ and $B^{'} = \{a_3, a_4\}$ then, $S_B \cap S_{B'} = \emptyset$. The set of attributes in $B^{'} - B = a_3$, represent the additional information required by objects in $S_{B'} = \{u_4\}$ to be distinguishable from the rest of objects.

2.2.6. Decision rules based on DORR

For each subgroup S_B, a set of decision rules can be generated following the syntax described in [15], as follows:

Decision rules generated from the B-lower approximation of the upward union of decision classes $Y_t^{\geq}$ are described by:

10

$$if \bigwedge_i (f(x,b_i) \geq r_{b_i}) \bigwedge \left(\bigwedge_j \left(f(x,a_j) = r_{a_j} \right) \right) then \ x \in Y_t^{\geq}$$

where $b_i \in B$ are criteria, $a_j \in B$ are attributes, $r_{b_i} \in V_{b_i}$ and $r_{a_j} \in V_{a_j}$. Decision rules generated from the B-lower approximation of the downward union of classes $Y_t^{\leq}$ are described by

$$if \bigwedge_i (f(x,b_i) \leq r_{b_i}) \bigwedge \left(\bigwedge_j \left(f(x,a_j) = r_{a_j} \right) \right) then \ x \in Y_t^{\leq}$$

where $b_i \in B$ are criteria, $a_j \in B$ are attributes, $r_{b_i} \in V_{b_i}$ and $r_{a_j} \in V_{a_j}$.

In our example, $Y^{\geq}$ and $Y^{\leq}$ correspond to the decision classes $Y_{Altered}$ and Y_{Normal}, respectively. Considering the subgroup $S_{a_4} = \{u_1, u_5\}$ and $S_{a_3,a_4} = \{u_4\}$, the following rules are induced:

For S_{a_4}:

$$\text{if } f(u, HoursSitting) \leq 0 \text{ then } u \in Y_{Normal} \qquad (u_1)$$

$$\text{if } f(u, HoursSitting) \geq 3 \text{ then } u \in Y_{Altered} \qquad (u_5)$$

For S_{a_3,a_4}

$$\text{if } f(u, Smoking) \geq 1 \text{ and } f(u, HoursSitting) \geq 2 \text{ then } x \in Y_{Altered} \qquad (u_4)$$

2.2.7. Comparing ORRs vs. DORRs

The DORRs allow identifying patients with inconsistent information and avoid the use of condition attributes that violates the dominance principle. The ORR method dismisses this important information and is reflected in the subgroups and decision rules obtained from the dataset.

For example, the information in patient u_3 suggests that his test results should be normal, yet the test results appears as altered. The inconsistent information appear in attributes a_2, a_3 and a_4.

As a result, the ORR grouping process results in patients u_1 and u_5 being in different groups. This suggests that u_1 needs additional information about either his smoking habits (a_2) or a disease presented during his childhood (a_3). However, after capturing the inconsistent attributes, the DORR's, assign both patient in the same group indicating that the only required information for both patients is the number hours they spend sitting per day (a_4).

Patients u_1, u_2 and u_4 appear in the same group under the OOR grouping scheme. On the other hand, the DORR's grouping process indicates that those patients belongs to different groups. Patient u_2, for example, needs different information than patient u_4. Patient u_2 requires information about his smoking habits (a_3) and the occurrence of a disease in their childhood (a_2).

The decision rules from the ORR grouping process contain inconsistencies in their descriptions as show in the rule induced form patient u_3. Moreover, to classify new cases, the condition part of the rule have to match exactly with the new patient attributes values. This disadvantage is diminished by the DORR decision rules as they include attribute value ranges in their descriptors.

2.3. Dataset description

The dataset used in this study consists of the 9105 cases from the SUPPORT (Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments) prognostic model dataset [18]. We consider as condition attributes the variables used in the SUPPORT prognostic model equation [19] to ensure consistency. The SUPPORT variables include ten physiologic variables in addition to the diagnosis groups, age, number of days in the hospital before entering the study, presence of cancer, and neurological function as recorded in the SUPPORT data. Attribute names, descriptions and value ranges are listed in Table 4.

Appendix F (continued)

Table 4: Description of attributes from SUPPORT dataset

Variable Name	Description
Numerical Condition Attributes	
age	Age of the patient
alb	Serum albumin
bili	Bilirubin
crea	Serum creatinine
hday	Number of days in hospital at study entry
hrt	Heart Rate
meanbp	Mean arterial blood pressure
pafi	Blood gasses, $PaO_2/(.01 * FiO2)$ 0
resp	Respiration rate
scoma	SUPPORT coma score, based on Glasgow coma scale
sod	Sodium
temp	Temperature in °C
wblc	White blood cell count
Categorical Condition Attributes	
dzgroup	Diagnosis Group: *ARF/MOSF w. Sepsis* CHF Cirrhosis Colon Cancer Coma COPD Lung Cancer MOSF w. Malignancy
ca	Presence of cancer: *Yes* *No* *Metastasis*
Decision Attribute	
d.6months	Death occurred within 6 months: *Yes* *No*

2.4. Data Preprocessing

To discretize the continuous variables we use the Acute Physiology and Chronic Health Evaluation (APACHE) III Scoring System [4], a clinically accepted scoring system designed to estimate the risk of death in ICU patients. APACHE III scores are designed to increase monotonically with respect to risk of death and thus provide the necessary preference relations for the DRSA. For further details see [16].

2.5. Performance Evaluation

The performance of our methodology is evaluated using a 5-fold cross validation procedure. In 5-fold cross validation, the entire dataset is randomly divided into five subsets, or folds, and then each fold (20% of the dataset) is used once as a testing set, with the remaining folds (80%) used for training. The performance results of the proposed methodology are compared with previous results obtained in [16], where we use the MODLEM and VC-DomLEM algorithms for inducing decision rules based on the classical and dominance-based rough set approaches. Both MODLEM and VC-DomLEM induce a minimum number of decision rules directly from the dataset to cover all objects from the lower approximations of the decision classes.

3. Results

Table 5, shows the performance comparison in terms of AUC and coverage for the three RST-based classification models. Our methodology, based on DORR subgroups, performs similar to the other RST-based the methodologies (Wilcoxon Signed-rank test $p - value > 0.05$).

To measure the accessibility of the model, we calculate the average number of rules that fire for classifying a new patient in the testing set. The CRSA with MODLEM algorithm generates the fewest number of rules fired per patient with a mean of 3.06 and standard deviation of 1.65;

Table 5: AUC and coverage between RST-based classifiers

	CRSA*		VC-DRSA**		DORR	
Fold	AUC (%)	Coverage (%)	AUC (%)	Coverage (%)	AUC (%)	Coverage (%)
1	69.96	97.38	71.54	97.97	67.85	99.06
2	70.10	97.47	72.53	98.57	68.25	98.75
3	67.58	97.64	70.48	98.96	67.53	97.23
4	71.78	97.58	73.64	98.40	69.54	98.32
5	69.21	96.98	70.44	98.45	67.21	97.45
Mean	69.73	97.41	71.73	98.47	68.08	98.16
Std. Dev.	1.52	0.26	1.37	0.36	0.90	0.80

* CRSA with MODLEM algorithm (α-consistency level = 0.6)
** VC-DRSA (l-consistency level = 0.6)

followed by the DORR method, showing a mean of 4.47 and a standard deviation equals to 8.74. Finally, the VC-DRSA has on average 13.65 rules that fire for each patient, with a standard deviation of 20.78.

The number of descriptors in a decision rule determine how general or specific is the rule. The MODLEM algorithm produces general rules with 3.65 descriptors on average and a maximum of 8 descriptors. The VC-DOMLEM decision rules are on average longer, more specific with mean and maximum length of 6.85 and 13 descriptors, respectively. The average length of the DORR rules is slightly lower, with rules containing on average 6.27 descriptors and a maximum of 15 descriptors.

4. Discussion

All three RST-based prognostic models perform comparable in terms of accuracy and accesibility by presenting the physician with a list of matched rules that offer significant advantages in terms of traceability of the model and the amount of information included in its results.

All the RST-based models analyze a clinical data set by exploring patients' characteristics. The MODLEM and VC-DOMLEM algorithms induce approximated decision rules to avoid overfitting the training set and to generate rules that are more useful in classifying new cases. This process also induces shorter rules with high support for their description [16]. However, rules with few attributes in their description can cause skepticism as some factors considered important in clinical practice may be omitted [20, 21]. Moreover, shorter rules are less likely to capture individual patient's characteristics necessary to develop a patient-centered model.

Table 6 shows the set of matched decision rules that classify the following example patient from the test set: A 52 years old patient with a primary diagnosis of coma and no cancer. The patient displayed hight head injury on the Glasgow Coma Scale; normal levels of creatinine, bilirrubin, albumin, temperature, and sodium; and moderated levels of respiratory rate (30 bpm), heart rate (120 bpm) and mean blood pressure (110 bpm). The patient survived 1728 days.

Both the DORR and VC-DRSA methods correctly predict that the patient will survive the six months period, however the MODLEM algorithm fails to classify the patient correctly. The MODLEM algorithm induces a very general rule with only one descriptor (rule 3 in Table 6). In the voting process, this rule practically decides the classification as it presents higher support to classify the patient. The other two more specific rules (rule 1 and 2 Table 6) have low support, yet correctly classify the patient. The VC-DOMLEM algorithm include rules with more specific information as shown in rule 4 and rule 5 from Table 6. These rules consider that the patient is relatively young with normal values in some of the physiological variables, such as temperature and sodium. This specific information gives the necessary support to correctly classify the patient. The proposed DORR-based method also induces patient-specific decision rules with the particular characteristic that all rules are deterministic, i.e. all cases matching the rule description support the classification with 100% accuracy. This characteristic removes the limitation of applying general information for classifying new patients as all cases matching the rule description fully support the classification of the patient. For example, rule 8 in Table 6 describes patients with a degree of coma less or equal to severe, with normal levels of albumin and white blood cell count less than 24.9 $cu/mm * 1000$. All the 10 patients matching these characteristics survived the six month period.

5. Conclusions

We introduce the definition of the DORR's to decompose the dataset into subgroups of patients with similar characteristics that differentiate them from the rest of patients with a different outcome. To classify new patients, our methodology determines the subgroups to which the patient belongs and use the decision rules corresponding to that subgroups only. This in contrast to common rule-based classifiers where the whole set of rules is used to classify new patients.

Our methodology performs similarly compared to the CRSA and VC-DRSA in terms of accuracy. The main advantage of the proposed DORR approach is that we achieved a higher coverage without using any approximation for generating the decision rules. We generate specific decision rules to classify a patient using minimal information as appears in its reducts.

The type of data analysis performed in this paper, open the opportunity for the knowledge extraction process to identifying groups of patients with similar characteristics allowing tailored decisions for the pertaining subgroup. Applications of this methodology could be useful to identify subgroups of patients that need different treatments, patients with differential response to therapy or patients that belong to different risk groups.

Appendix F (continued)

Table 6: Decision rules fired from CRSA-MODLEM and VC-DRSA

			RHS Support	
	CRSA-MODLEM Rules	LHS	*d.6months = No*	*d.6months = Yes*
1.	If *resp_score*[a] =6 AND *age_score*[b]=5 AND *meanbp_score*[c]=4	98	62 (63.27%)	36 (36.73%)
2.	If *dzgroup*=Coma AND *age_score*[b]=5 AND *hrt_score*[d]=7 AND *meanbp_score*[c]=4 AND *crea_score*[e]=0	4	3 (75%)	1 (25%)
3.	If *scoma* =Severe	517	79 (15.28%)	438 (84.72%)
	VC-DRSA Rules			
4.	If *dzgroup*=Coma AND *scoma*≤ *S evere* AND *age_score*[b] ≤ 5 AND *temp_score*[f] ≤ 0	15	11 (73%)	4 (27%)
5.	If *dzgroup*=Coma AND *scoma*≤ *S evere* AND *age_score*[b]≤ 11 AND *temp_score*[f] ≤ 0 AND *sod_score*[g]≤ 0	12	11 (85%)	2 (15%)
	DORR Rules			
6.	If *dzgroup*=Coma AND *wbc_score*[h]≤ 1 AND *alb_score*[i] ≤ 0	2	2 (100%)	0 (0%)
7.	If *dzgroup*=Coma AND *alb_score*[i]≤ 0 AND *crea_ score*[e]≤ 4	7	7 (100%)	0 (0%)
8.	If *Scoma*≤ *S evere* AND *wbc_score*[h]≤ 1 AND *alb_score*[i]≤ 0	10	10 (100%)	0 (0%)

[a] *resp_score:* 6 = (25 ≤ *resp* ≤ 34)

[b] *age_score:* 5 = (45 ≤ *age* ≤ 59) ; ≤ 5 = (*age* ≤ 59) ; ≤ 11 = (*age* ≤ 64)

[c] *meanbp_score:* 4 = (100 ≤ *meanbp* ≤ 119)

[d] *hrt_score:* 7 = (120 ≤ *hrt* ≤ 139)

[e] *crea_score:* 0 = (0.5 ≤ *crea* ≤ 1.4) ; 4 = (1.5 ≤ *crea* ≤ 1.94)

[f] *temp_score:* ≤ 0 = (36 ≤ *temp* ≤ 36.9)

[g] *sod_score:* ≤ 0 = (135 ≤ *sod* ≤ 154)

[h] *wbc_score:* ≤ 1 = (*wbc* ≤ 24.9)

[i] *alb_score:* ≤ 0 = (2.5 ≤ *alb* ≤ 4.4)

6. References

References

[1] A. R. Claxton R, Angus D, Prognostic Models in Critically Ill Patients. Fast Facts and Concepts, 2010.

[2] W. J. Ehlenbach, C. R. Cooke, Making ICU prognostication patient centered: is there a role for dynamic information?, Critical care medicine 41 (4) (2013) 1136–8.

[3] M. Maltoni, A. Caraceni, Brunelli, et al., Prognostic factors in advanced cancer patients: evidence-based clinical recommendations, A study by the Steering Committee of the European Association for Palliative Care, Journal of clinical oncology : official journal of the American Society of Clinical Oncology 23 (25) (2005) 6240–8.

[4] W. A. Knaus, D. P. Wagner, Draper, et al., The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults, Chest 100 (6) (1991) 1619–1636.

[5] J. Rogers, H. D. Fuller, Use of daily Acute Physiology and Chronic Health Evaluation (APACHE) II scores to predict individual patient survival rate, Critical care medicine 22 (9) (1994) 1402–5.

[6] J. M. Teno, J. E. Shu, et al., Timing of Referral to Hospice and Quality of Care: Length of Stay and Bereaved Family Members' Perceptions of the Timing of Hospice Referral, Journal of Pain and Symptom Management 34 (2) (2007) 120–125.

[7] M. Kashani-Sabet, R. W. Sagebiel, Joensuu, et al., A Patient-Centered Methodology That Improves the Accuracy of Prognostic Predictions in Cancer, PLoS ONE 8 (2) (2013) e56435.

[8] I. R. Konig, J. D. Malley, S. Pajevic, et al., Patient-centered yes/no prognosis using learning machines, Int J Data Min Bioinform 2 (4) (2008) 289–341.

[9] S. Visweswaran, D. C. Angus, M. Hsieh, others., Learning patient-specific predictive models from clinical data, Journal of biomedical informatics 43 (5) (2010) 669–85.

[10] C. Gottrup, K. Thomsen, P. Locht, et al., Applying instance-based techniques to prediction of final outcome in acute stroke, Artificial Intelligence in Medicine 33 (3) (2005) 223–236.

[11] I. of Medicine, Media Reviews, Journal for Healthcare Quality 24 (5) (2002) 52–54, ISSN 1945-1474.

[12] J. C. Wyatt, D. G. Altman, Commentary: Prognostic models: clinically useful or quickly forgotten?, BMJ 311 (7019) (1995) 1539–1541.

[13] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning about Data, Norwell, MA, 1992.

[14] J. Stepaniuk, Approximation Spaces, Reducts and Representatives, 1998.

[15] J. Błaszczyński, S. Greco, R. Słowiński, Multi-criteria classification–A new scheme for application of dominance-based decision rules, European Journal of Operational Research 181 (3) (2007) 1030–1044.

[16] E. Gil-Herrera, G. Aden-Buie, A. Yalcin, et al., Rough Set Theory based Prognostic Model for Hospice Referral, Artificial Intelligence of Medicine (in second round review) .

[17] Y. Yao, Y. Zhao, Discernibility matrix simplification for constructing attribute reducts, Information Sciences 179 (7) (2009) 867–882.

[18] F. E. Harrell, SUPPORT Datasets, http://biostat.mc.vanderbilt.edu/wiki/Main/SupportDesc, 2010.

[19] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, D. P. Wagner, The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults, Annals of Internal Medicine 122 (3) (1995) 191–203.

[20] E. Steyerberg, Clinical Usefulness, in: Clinical Prediction Models, Statistics for Biology and Health, Springer New York, ISBN 978-0-387-77243-1, 281–297, 2009.

[21] M. Ebell, AHRQ White Paper: Use of clinical decision rules for point-of-care decision support, Med Decis Making 30 (6) (2010) 712–21.

[22] F. C. Brosius, T. H. Hostetter, E. Kelepouris, et al., Detection of chronic kidney disease in patients with or at increased risk of cardiovascular disease: A science advisory from the American Heart Association Kidney And Cardiovascular Disease Council, Circulation 114 (10) (2006) 1083–7.

About the Author

Eleazar Gil-Herrera received his B.S. in Computer Engineering from University of San Antonio Abad in Cusco, Peru. In 2009, he received his M.S in Industrial and Systems Engineering from the University of Puerto Rico at Mayaguez. He received his Ph.D. in Industrial and Management Systems Engineering from the University of South Florida in 2013. His areas of research include clinical data analytics, development of prognostic models and decision support models for clinical decision-making.