



UNIVERSIDAD REY JUAN CARLOS

TESIS DOCTORAL

MACHINE LEARNING AND KNOWLEDGE MANAGEMENT

FOR DECISION SUPPORT. APPLICATIONS IN

PROMOTIONAL EFFICIENCY AND HEALTHCARE

Autora:

Cristina Soguero Ruiz

Directores:

Dr. José Luis Rojo Álvarez

Dra. Inmaculada Mora Jiménez

DEPARTAMENTO DE TEORÍA DE LA SEÑAL Y COMUNICACIONES
Y SISTEMAS TELEMÁTICOS Y COMPUTACIÓN

Fuenlabrada, mayo 2015

José Luis Rojo Álvarez, con D.N.I. 09788715-F, e Inmaculada Mora Jiménez con D.N.I. 04591755N, como Directores de la Tesis Doctoral realizada por Cristina Soguero Ruiz y titulada *Machine Learning and Knowledge Management for Decision Support. Applications in Promotional Efficiency and Healthcare*, hacemos constar que ésta cumple con todos los requisitos necesarios, y por ello autorizamos la defensa de la misma.

Fdo.

Fuenlabrada, a de de 2015

TESIS DOCTORAL

**Machine Learning and Knowledge Management for Decision Support.
Applications in Promotional Efficiency and Healthcare**

AUTOR: *CRISTINA SOGUERO RUIZ*

DIRECTORES: *DR. JOSÉ LUIS ROJO ÁLVAREZ*

DRA. INMACULADA MORA JIMÉNEZ

Firma del Tribunal Calificador:

Firma

PRESIDENTE:

VOCAL:

VOCAL:

VOCAL:

SECRETARIO:

Calificación:

Fuenlabrada, a de de 2015

Resumen

El desarrollo alcanzado en las Tecnologías de la Información y las Comunicaciones en las últimas décadas, ha traído consigo la recopilación y almacenamiento creciente de datos en ámbitos tan diversos como pueden ser marketing, salud o seguridad. La disponibilidad de grandes cantidades de datos hace necesaria la búsqueda de nuevos paradigmas de aprendizaje máquina, capaces de abordar el análisis automatizado de los mismos con la consiguiente extracción de información.

En concreto, las técnicas de *aprendizaje máquina* permiten diseñar modelos estadísticos no paramétricos que aprendan las relaciones existentes entre un conjunto suficientemente representativo de ejemplos, cada uno de ellos formado por unas variables observadas (características), y su correspondiente salida. Se desea que el modelo construido pueda generalizar, es decir, obtener una salida adecuada ante ejemplos de entrada no considerados durante la fase del diseño. En los últimos años, estas técnicas han experimentado un avance espectacular, tanto en fundamentos teóricos como en su aplicación a distintos y numerosos dominios de conocimiento.

El *objetivo general* de esta Tesis es el desarrollo teórico y la implementación de métodos de aprendizaje máquina, con énfasis en las etapas de selección de características y diseño del modelo predictivo, de forma que permita abordar el análisis de grandes cantidades de datos de naturaleza diversa, creando procedimientos específicos para cada etapa pero al tiempo aplicables en distintos ámbitos.

En esta Tesis se han abordado tres áreas específicas de creciente interés económico y social: (a) el modelado de las interacciones entre productos de consumo diario y su eficiencia promocional; (b) el apoyo a la toma de decisiones para la predicción temprana de complicaciones tras la cirugía de cáncer de colon; (c) la estratificación de riesgo de muerte súbita cardíaca a partir de índices predictores obtenidos de las señales eléctricas del corazón, utilizando un modelo de conocimiento clínico y una terminología estandarizada. El análisis de datos de cada una de estas aplicaciones presenta como denominador común la utilización de técnicas de aprendizaje

máquina, de acuerdo con el objetivo general. Sin embargo, la naturaleza tan diversa de dichas aplicaciones hace que cada una represente por sí misma un objetivo específico de la presente Tesis.

El *primer objetivo específico* consiste en profundizar en la evaluación y análisis de las ventas promocionales, tradicionalmente basado en técnicas de estadística clásica. Un apoyo sustancial en la toma de decisiones ha de venir necesariamente del análisis sistemático de datos masivos sobre el control y monitorización de las promociones y sus complejas interacciones. Por ello se propone el análisis y la comparación estadística de distintas técnicas de aprendizaje máquina.

Otro ámbito de naturaleza muy diversa al anterior, pero de indudable interés social, es el de la salud. El análisis de datos clínicos, tanto estructurados (constantes vitales o análisis de sangre) como no estructurados (texto libre en documentos clínicos), recogidos longitudinal y sistemáticamente en las historias clínicas electrónicas (HCEs) de un conjunto numeroso de pacientes, permite incrementar sustancialmente el conocimiento clínico y apoyar la toma de decisiones. Sin embargo, las técnicas de aprendizaje máquina y el análisis de datos han tenido, hasta la fecha, un alcance limitado en este ámbito. Esta situación se debe principalmente a la dificultad de extraer información útil de datos clínicos procedentes de fuentes heterogéneas. Además, existen muy pocos precedentes de sistemas que permitan la explotación automática de la información a nivel agregado entre diferentes entidades hospitalarias y existe gran necesidad de disponer de datos que sirvan de base para el avance científico, con mayor impacto en la práctica clínica. En esta Tesis se analizan dos dominios del ámbito salud de gran prevalencia en el mundo occidental, a saber, el cáncer de colon y las enfermedades cardíacas.

El *segundo objetivo específico* consiste en la adaptación y aplicación de métodos de aprendizaje máquina para la detección temprana de complicaciones tras la cirugía de cáncer de colon, analizando tanto individual como conjuntamente variables procedentes de fuentes heterogéneas, extraídas todas ellas de la HCE.

El *tercer objetivo específico* consiste en la creación de modelos de conocimiento clínico que permitan intercambiar datos y comprender semánticamente la información clínica de distintas HCEs. En los últimos años se han propuesto numerosos índices predictores del riesgo cardíaco. En concreto, en esta Tesis se analiza el dominio de la turbulencia del ciclo cardíaco por ser un predictor de muerte súbita cardíaca con guías clínicas claras y concisas.

El análisis de grandes cantidades de datos y el desarrollo teórico de nuevos algoritmos de aprendizaje estadístico representan hoy, sin duda, un área de investigación muy activa en distintos dominios. Esta Tesis contribuye a mejorar el conocimiento y la toma de decisiones en aplicaciones reales de muy diversa naturaleza, y al tiempo con claros denominadores comunes.

Abstract

The development achieved in Information and Communications Technologies in recent decades has brought an enormous growth in the collection and storage of data in such diverse fields as marketing, health, or safety. The availability of large amounts of data makes necessary the search for new machine learning paradigms, capable of addressing their automated analysis and the subsequent information extraction. Specifically, given a number of training examples (also called samples or observations) associated with desired outcomes, the *machine learning* techniques learn the relationship between them. In recent years, these techniques have experienced spectacular advances in both theoretical foundations and their application to a wide range of different knowledge domains.

The *general objective* of this Thesis consists on the theoretical development and implementation of machine learning methods, with emphasis on the feature selection and predictive model design stages, allowing to tackle with the analysis of data of diverse nature, and creating specific procedures for each stage but at the same time applicable in various fields.

This Thesis has addressed three specific areas of increasing economic and social interest: (a) interaction modeling between everyday products and promotional efficiency; (b) clinical decision support for early detection of complications after colorectal cancer surgery; (c) risk stratification of sudden cardiac death from predictive indices obtained from the electrical signals of the heart, using a clinical knowledge model and a standardized terminology. The data analysis in these applications shares the use of machine learning techniques according to the general goal. However, the diverse nature of these applications represents by itself a specific goal of this Dissertation.

The *first specific objective* consists on further evaluation and analysis of promotional sales, traditionally based on classical statistical techniques. A substantial support decision making must necessarily come from the systematic analysis of massive data on the control and monitoring of promotions and their complex interactions. Therefore, a statistical analysis and comparison of various machine learning techniques is proposed.

Another area of very different nature respect to the previous one, but with strong social interest, is healthcare. The analysis of clinical data, both structured (vital signs or blood tests) and unstructured (text-based documents), systematically and longitudinally collected from the electronic health record (EHR) of a large group of patients, can substantially increase the clinical knowledge and support decision-making. However, machine learning techniques and massive data analysis have provided, nowadays, a limited impact in the healthcare area. This situation is mainly due to the difficulty of extracting useful information from clinical data recorded in heterogeneous sources. In addition, there are few precedents of systems enabling the automatic analysis of information at the aggregated level among different hospital entities. There is a great need for suitable and relevant data as a basis for scientific advance, with greater impact on the clinical practice. In this Thesis, two healthcare domains highly relevant in most developing countries are analyzed, namely, colorectal cancer and cardiovascular diseases.

The *second specific objective* is the adaptation and application of machine learning methods for early detection of complications after colorectal cancer surgery, analyzing both individually and jointly data from heterogeneous sources recorded in the EHR.

The *third specific objective* is to build clinical knowledge models to enable data exchange and semantical understanding of clinical information from different EHR. In recent years, numerous predictors of cardiac risk indices have been proposed. Specifically, in this Thesis, the heart rate turbulence is analyzed to be a predictor of sudden cardiac death with clear and concise guidelines.

Nowadays, the analysis of large amounts of data as well as the theoretical development of new machine learning algorithms undoubtedly represent a very active area of research in different domains. This Thesis contributes to improve knowledge and decision making in real-world applications of diverse nature which still share remarkable common denominators.

Agradecimientos

Just don't give up on trying to do what you really want to do. Where there is love and inspiration, I don't think you can go wrong.

–Ella Fitzgerald

La culminación de esta Tesis ¹ ha sido posible gracias a la colaboración de todas las personas y entidades que han confiado en ella y han hecho posible que llegue a su fin. Quisiera agradecerles a todos su confianza y dedicarles este trabajo.

En primer lugar me gustaría dar las GRACIAS a mis directores José Luis Rojo Álvarez e Inmaculada Mora Jiménez, y no solo por su papel como directores, para mí son mucho más que eso. El destino quiso que nos encontrásemos hace ya algunos años, y desde ese momento me han brindado la oportunidad de aprender y de crecer junto a ellos. No tengo palabras para agradecerles todo el tiempo, esfuerzo y apoyo que me han ofrecido y, que sin duda, ha sido el pilar base de esta Tesis. Espero que el destino nos siga uniendo muchos años más. GRACIAS!

Mi especial agradecimiento a Robert Jenssen, por la confianza depositada en mí, y las oportunidades que me ha facilitado. Mis estancias en el Polo Norte, me han permitido crecer tanto profesionalmente como personalmente, y su apoyo ha sido fundamental. THANK YOU!

A todos los miembros que forman y han formado parte del Departamento de Teoría de la Señal y Comunicaciones y Sistemas Telemáticos y Computación de la Universidad Rey Juan Carlos, por su colaboración en la realización de esta Tesis, por las charlas de investigación, por el día a día, por los cafés, por las risas, por la ayuda prestada. Gracias a todos! Sin duda, habéis hecho que etapa sea mucho más agradable. A Javier Ramos, por su papel como director de la Escuela, por su constante interés por los temas de esta Tesis.

A todas las entidades y miembros de las mismas, que sin duda alguna, han permitido poner el punto final a esta etapa. En primer lugar a la Universidad Rey Juan Carlos, la cual ha sido partícipe de toda mi carrera universitaria. A A.T. Kearney, a Natxo, a Carlos, por permitirme conocer qué significa trabajar con datos reales. A María Pilar Martínez y a Javier Gimeno, por enseñarme conceptos relacionados con marketing y con el mundo empresarial. Al Hospital Universitario de Fuenlabrada, especial mención a Luis Lechuga, por acercarme al mundo del

¹Cristina Soguero Ruiz es becaria del programa de Formación de Profesorado Universitario (FPU) con referencia AP2012-4225.

Este trabajo ha sido parcialmente financiado por el proyecto de investigación TEC2010-19263 del Ministerio de Ciencia e Innovación.

modelado del conocimiento clínico. Al Hospital Virgen de la Arrixaca de Murcia, a Arcadi. A la Universidad de Tromso, a Jonas, a Karl Øyvind, a Muhammad, por hacer mis estancias más acogedoras y entretenidas. Al Hospital Universitario de Tromso y al Centro Noruego para la Atención Integral y Telemedicina, a Stein, a Knut Magne, a Fred, a Kristian, a los cirujanos, por facilitarme datos clínicos reales, y por su colaboración en el análisis de los mismos. A Persei Consulting, a Roberto Bravo. A mis ex-compaños de Telefónica I+D, a Enrique, a Vanessa, a Víctor, a Jesús, a Giovanna.

A mis amigos, por su apoyo, por aguantar ese repetido no. Gracias. Habéis sido una parte fundamental de esta Tesis. A Anita, por sus consejos, por estar a mi lado cada día, a Elena, a Javi, a Marta, a mis chicas, a los hidalgos.... a Gema, por su cariño y apoyo constante, especialmente cuando más lo necesitaba.

A Juan, por su apoyo incondicional, por estar siempre a mi lado, por aguantar mis épocas de estrés, de alegría, de tristeza. GRACIAS por formar parte de este proyecto.

Y por supuesto, no puede faltar mi familia. GRACIAS por todo el cariño y todo el apoyo que me han dado cada día. En especial a mis padres, siempre han sido el motor que he necesito para caminar. A mi hermana, por estar siempre ahí sin recibir nada a cambio. a mi sobrino, por sus sonrisas y su infinita alegría en los momentos que más lo necesitaba. A mi abuela.

Cristina Soguero Ruiz

Contents

1	Introduction	19
1.1	Background and Motivation	19
1.2	General and Specific Objectives	22
1.3	Thesis Structure and Contributions	23
2	Fundamentals and Contributions in Machine Learning	27
2.1	Introduction	27
2.2	Feature Engineering	28
2.2.1	Background	28
2.2.2	Contribution 1. Statistical Characterization of Features	30
2.3	Predictive Modeling	31
2.3.1	Machine Learning Methods	31
2.3.2	Contribution 2. Covariance Kernel Series for Regression	37
2.4	Feature Selection	41
2.4.1	Conventional Methods	41
2.4.2	Contribution 3. Statistical Feature Selection Strategies	42
2.5	Model Selection	46
2.5.1	Merit Figures and Generalization Evaluation	46
2.5.2	Contribution 4. Bootstrap Resampling for Benchmarking Machine Learning Models	48
3	Machine Learning for Promotional Decision-Making	53
3.1	Introduction	53
3.2	Application 3.1: Promotional Efficiency at Store Level	55
3.2.1	Introduction	55
3.2.2	Database	57

3.2.3	Experiments and Results	59
3.2.4	Discussion and Conclusions	66
3.3	Application 3.2: Promotional Efficiency at Chain Level	67
3.3.1	Introduction	67
3.3.2	General Forecasting Promotional Model	69
3.3.3	Database	71
3.3.4	Experiments and Results	74
3.3.5	Discussion and Conclusions	81
4	Machine Learning for Healthcare Analytics	83
4.1	Introduction	83
4.2	Application 4.1: Early Detection of Anastomosis Leakage from Bag-of-Words	84
4.2.1	Introduction	84
4.2.2	Database	85
4.2.3	Experiments and Results	86
4.2.4	Discussion and Conclusions	93
4.3	Application 4.2: Early Detection of Anastomosis Leakage using Heterogeneous Sources	95
4.3.1	Introduction	95
4.3.2	Database	96
4.3.3	Experiments and Results	98
4.3.4	Discussion and Conclusions	103
4.4	Application 4.3: Data-driven Temporal Prediction of Surgical Site Infection	104
4.4.1	Introduction	104
4.4.2	Database	105
4.4.3	Experiments and Results	107
4.4.4	Discussion and Conclusions	110
5	Knowledge Management in Electronic Health Record	113
5.1	Introduction	113
5.1.1	Heart Rate Turbulence	114
5.1.2	The Conceptual Model of SNOMED-CT	116
5.1.3	CEN/ISO EN13606 standard	118
5.2	Application 5.1: Ontology for Clinical Decision Support in EHR	119
5.2.1	Ontology based on SNOMED-CT	121
5.2.2	Ontology Prototype in EHR	122
5.2.3	HRT Clinical Decision Support	125

5.3	Application 5.2: From Archetypes to EHR Web Prototype	126
5.3.1	HRT Archetype Prototype	127
5.3.2	Server-Based Ontology System Archetype Binding	129
5.3.3	Clinical Data Export for Semantic Interoperability	130
5.4	Discussion and Conclusions	132
6	Conclusions and Future Work	135
6.1	Conclusions	135
6.2	Future Work	137

List of Acronyms and Abbreviations

ADL	Archetype Definition Language
AL	Anastomosis Leakage
ALAT	Alanine Aminotransferase
ALP	Alkaline Phosphatase
ANN	Artificial Neural Networks
ARMA	Auto-Regressive Moving-Average
BL	Base Line
BoW	Bag-of-Words
CDS	Clinical Decision Support
CI	Confidence Interval
CM	Confusion Matrix
CRC	Colorectal Cancer
CRP	C-Reactive Protein
CVRS	Cardiovascular Risk Stratification
DD	Direct Discount
DEC	Deal Effect Curve
EHR	Electronic Health Record
FDA	Fisher's Discriminant Analysis
FMA	Foundational Model of Anatomy Ontology
FS	Feature Selection
GP	Gaussian Processes
GRNN	General Regression Neural Network
HIS	Hospital Information Systems
HQ	Headquarters
HRT	Heart Rate Turbulence
HRV	Heart Rate Variability
ICD10	International Classification of Diseases
IHTSDO	International Health Terminology Standards Development Organization

KDE	Kernel Density Estimation
KECA	Kernel Entropy Component Analysis
<i>k</i> -NN	<i>k</i> Nearest-Neighbors
LOCF	Last Observation Carried Forward
LOO	Leave One-Out
MAE	Mean Absolute Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
NB	Naive Bayes
NCSP	NOMESCO Classification of Surgical Procedures
PI	Price Indices
RBF	Radial Basis Function
RFE	Recursive Feature Elimination
RKHS	Reproducing Kernel Hilbert Space
SWRL	Semantic Web Rule Language
SNOMED-CT	Systematized Nomenclature of Medicine - Clinical Terms
SSI	Surgical Site Infections
SVM	Support Vector Machine
TS	Time Series
TWA	T-wave Alternans
UMLS	Unified Medical Language System
VPC	Ventricular Premature Complex

Introduction

1.1 Background and Motivation

Over recent years, there has been an enormous growth in available data which is getting ever vaster and ever more rapidly in a wide range of different fields. Analyzing data allows us to obtain knowledge and support decision-making in a number of real-world applications. Machine Learning (ML) methods [1, 2, 3, 4] have been proposed as key tools to lead new breakthroughs that will improve the human abilities for analyzing many data types. ML techniques allows to learn the relationships among a number of input training samples (observation or examples) and a desired output [4, 5, 6], each sample being described by a number of binary, continuous, or categorical variables (features). The goal of the learning process is to predict the outcome value for a new sample (test sample), and a predictive model is built towards that end.

The elements of the predictive modeling pipeline, as shown in Fig. 1.1, are feature extraction, model design, and prediction model [7]. Feature extraction is very domain specific, and expert knowledge is required to come up with a useful number of features. Sometimes, raw features are directly used as input variables in the model design whereas, in other cases, features are built from the original variables after a preprocessing or engineering stage. This corresponds to the first element of the model design, so-called *feature engineering*. The second element, *feature selection*, is primarily performed to select relevant and informative features [8]. The main idea of the *model exploration* consists on choosing a mathematical method for prediction of the desired outputs from a set of variables [4]. Once the model is developed, expert interpretation and useful conclusions obtained from the prediction model are needed to support decision-making in real-world applications.

Despite the large amount of theoretical work developed on the previously described elements, there is no universal statistical framework to be used in all the applications. Hence, many design decisions are taken either heuristically, or guided by the vast experience of the ML

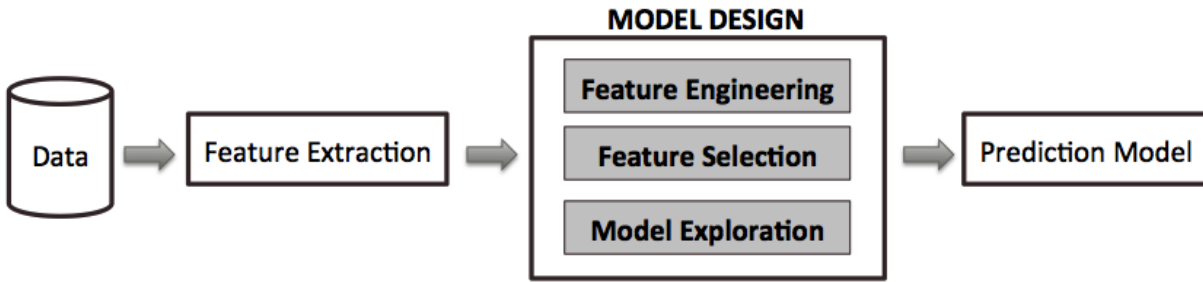


Figure 1.1: *Predictive modeling pipeline.*

systems, or founded on complex statistical backbones for specific elements. Therefore, it would be highly desirable a statistical framework for decision-making in these elements during the design stage, allowing the expert to make design decision on clear or at least operative cut-off tests, but supported by statistical well-founded principles keeping easy to use. In this scenario, nonparametric statistical approaches can be used because they do not assume any restriction associated with the data distribution, only considering general assumptions about the nature of the distribution. The use of nonparametric statistics to modify the previous stages, as well as to propose new ML techniques, can be an operative, useful, and adaptable methodology to work with.

On the other hand, ML methods have been applied in a large number of practical areas such as neuroscience, bioinformatics, intelligent systems, finance, or behavioral targeting [3, 4]. Examples of specific applications are optical character recognition [9], speech recognition [10, 11], or web page ranking [12]. In this Dissertation, ML methods are applied in two separate domains of great interest nowadays, namely, promotional efficiency and healthcare.

Marketing and sales have been some of the most active applications of statistical learning [3, 13], due to the recent increased technology capabilities to store huge amounts of customer information [14]. ML techniques aim to find recurrent patterns, trends, or rules, which can explain the data behavior in a given context, and then allow the user to extract new knowledge on the consumer behavior, to improve the performance of marketing operations [15, 16, 17]. Empirical models for analyzing sales promotions effects have been used in the literature. However, more recent works focus on ML techniques as powerful tools to extract information from existing recorded data [18, 19]. A vast amount of knowledge has been extracted using ML techniques, although not all the promotional behaviors have been definitely studied and there is still room for deep and further analysis [15, 16, 17]. More specifically, operational problems arise in ML promotional modeling for evaluating their working hypothesis [19, 20, 21, 22, 23, 24, 25]. The use of conventional parametric tests are often not appropriate due to the heavy tails and

heteroscedasticity for the prediction residuals are often no longer supported.

A very different nature domain, but with undoubtedly increasing importance, is healthcare. Healthcare analytics are based on data extracted from Electronic Health Records (EHRs), which are collections of health information in digital storage format conveying the relevant information of a patient [26], and they contain routinely amassed quantitative data (e.g., laboratory tests), qualitative data (e.g., text-based documents), and transactional data (e.g., records of medication delivery). A considerable amount of literature exists on knowledge extraction from the EHRs, aimed to support clinical decision-making in several domains [27, 28, 29, 30, 31, 32, 33]. In this Thesis, we focus on two relevant clinical problems, namely, colorectal cancer (CRC) and cardiovascular diseases. On the one hand, according to the American Cancer Society, colorectal cancer is the third most common cancer in men and women in developed countries, and the third leading cause of cancer death. On the other hand, according to World Health Organization, cardiovascular diseases are the leading causes of death worldwide.

Recent studies shown that surgery is the only curative treatment for CRC [34]. However, standard elective colorectal resection is usually associated with a complication rate of 20-30%, which has severe implications for the patient [35]. Anastomosis leakage (AL) is among the most dreaded complications after CRC surgery. It is reported to occur in 5-15% of all patients who underwent colorectal cancer surgery, and it is recognized as an important quality indicator of the surgery procedure [36]. AL may be a lethal condition, therefore its early detection is vital [36, 37]. Authors in [38] showed that the risk of AL as determined by surgeons' risk assessment appeared to have low predictive value. A colon leakage score was developed in [39] to predict the risk of AL based on information from the literature and experts opinions, showing that this score is a good predictor for AL. However, novel methods to identify and detect this complication at an early stage are needed, specially to deal with the common heterogeneity and sparsity of EHR data.

On the other hand, advanced data processing methods that extract useful diagnostic information often do not reach the medical practice and that research effort does not benefit the society. For example, ML techniques and massive data analysis have had, to date, a limited impact in healthcare. This situation is due to the difficulty of extracting useful information from heterogeneous clinical sources that are not easy to process jointly. In addition, there are very few precedents of actual systems that allow the exploitation of aggregated information from different hospital entities. The use of standards aims to allow the interoperability among different systems, in order to provide to citizens and professionals with the access to clinical information anywhere. The definition of clear and standardized connections among the current scientific knowledge, its availability for the care community, and the actual patient databases, are becoming fundamental needs for the clinical practice.

Cardiovascular risk stratification (CVRS) is a key element to raise population awareness of diseases causing a significant burden of morbidity and mortality, as well as to identify and assess the correct diagnosis and therapy [40]. A wide variety of indices, such as heart rate variability (HRV) or T-wave alternants presence, can be obtained from the electrocardiogram (ECG) recordings and can be used as cardiovascular risk predictors, but they are not established in the clinical practice yet. To overcome this situation, standards and clinical knowledge management tools are required to achieve the interoperability in this domain.

1.2 General and Specific Objectives

The *general objective of this Thesis* is to develop new tools to adjust the predictive modeling pipeline to real-world data characterized by high dimensionality, sparsity, temporal dynamics, and scarcity in the number of samples. Specifically, a nonparametric feature engineering technique, a smoothing regression method based on covariance properties, three different feature selection (FS) strategies, and a methodology to benchmark predictive models, are proposed.

These theoretical contributions are applied in three different domains, thus, *three specific objectives* are defined:

1. To perform a novel data-driven approach to characterize promotional efficiency at both store and chain level. The new economic conditions have led to innovations in retail industries, such as more dynamic retail approaches based on flexible strategies. The assessment of promotional sales with models constructed by ML techniques can be readily used for agile decision-making. A reliable quantification tool is proposed in this work as an effective information system leveraged on recent and historical data that provides the managers with an operative vision.
2. To infer new knowledge from complex heterogenous patient longitudinal records for supporting the early detection of several complications after CRC surgery. In this Dissertation, unstructured (text-based documents) and structured data (laboratory tests and clinical signs) are extracted from EHR and analyzed to improve clinical outcomes and detect post-surgery complications at an early stage. ML techniques are used to deal with the sparsity and irregular sampling presented in this kind of data, as well as to build the predictive models.
3. To open the road towards achieving the interoperability in EHR data exchange and follow-up, the standardization of CVRS based on heart rate turbulence (HRT) domain is considered as a first step according to its clear and well-defined guidelines. Towards

that end, a prototype based on clinical knowledge modeling tools is built to enable the interoperability of HRT domain as well as the continuous improvement and research on it.

1.3 Thesis Structure and Contributions

The remainder of this Dissertation starts dealing with the general objective, in which an introduction to ML theory, as well as the proposed theoretical contributions, are presented. The three following chapters present in detail the specific objectives addressed for each domain. In the last chapter, conclusions and future lines are discussed.

In Chapter 2, the second step (model design) of the predictive modeling pipeline shown in Fig. 1.1 is presented. An overview of the state of the art, as well as the theoretical contributions in each stage, are described. Regarding feature engineering, a nonparametric technique is proposed: (1) to describe the individual behavior of each feature; and (2) to estimate the statistical distribution of the output conditioned to the input features. A new smoothing regression method based on the properties of the covariance matrix, called *Covariance Kernel Series*, is proposed. Later, the three proposed FS strategies in this Dissertation are explained. Figures of merit and generalization evaluation are explained for both classification and regression methods. Finally, a strategy based on nonparametric bootstrap resampling approach is developed to benchmark prediction models.

Chapter 3 presents the first application, whose objective is the development of a data-driven model to characterize promotional efficiency at store and chain levels for different product categories. The proposed method is based on ML techniques, as a useful way to analyze the multiple and simultaneous effects coexisting in promotional activities in retail markets when using real-world data. Different ML methods are analyzed and benchmarked by using an operative and simple statistical method based on bootstrap resampling proposed in this Thesis.

Chapter 4 presents the healthcare analytics applied for early detection of complications after CRC surgery, and for predicting surgical site infections at both pre-operative and post-operative stages. Towards that end, heterogeneous structured (laboratory tests and vital signs) and unstructured (text-based documents) data from the EHR are individually and jointly analyzed. Clinical data are sparse, high dimensional, scarce in terms of number of samples and time-dependent, which represent several challenges to deal with. First, clinical notes extracted from the EHR are used for early detection of complications after CRC surgery. It is informative to know whether the discriminatory power for identifying unhealthy patients increases when heterogeneous sources, such as laboratory tests and vital signs, are considered in an incremental way. Finally, different methods to deal with sparsity are benchmarked.

Chapter 5 presents the contribution related to knowledge management in EHR for the CVRS

domain. The design and use of a standard clinical terminology and a clinical knowledge model are proposed to get a standardized tool for clinical decision support, providing with technically straightforward inclusion of the HRT domain in the EHR. In addition, a web prototype is built in order to support HRT recordings allowing a simple follow-up by the medical community.

Chapter 6 is devoted to general conclusions and future work.

Since this Thesis presents a multidisciplinary work with a combination of both theoretical and practical approaches, specific introduction as well as topic devoted conclusions are presented for each application.

Contributions

The work developed in this Dissertation has been published on journal articles with impact factor, as well as presented in international conferences, which are next enumerated.

Journal articles

- [1] C. Soguero-Ruiz, F. J. Gimeno-Blanes, I. Mora-Jiménez, M. P. Martínez-Ruiz, J. L. Rojo-Álvarez, “On the differential benchmarking of promotional efficiency with machine learning modeling (I): Principles and statistical comparison”, *Expert Systems with Applications*, vol. 39, no. 17, pp. 12772-12783, 2012.
- [2] C. Soguero-Ruiz, FJ. Gimeno-Blanes, I. Mora-Jiménez, M. P. Martínez-Ruiz, J. L. Rojo-Álvarez, “On the differential benchmarking of promotional efficiency with machine learning modeling (II): Practical applications”, *Expert Systems with Applications*, vol. 39, no. 17, pp. 12784-12798, 2012.
- [3] C. Soguero-Ruiz, L. Lechuga, I. Mora-Jiménez, J. Ramos-López, A. García-Alberola and J. L. Rojo-Álvarez. “Ontology for heart rate turbulence domain from the conceptual model of SNOMED-CT”. *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1825-1833, 2013.
- [4] C. Soguero-Ruiz, F. J. Gimeno-Blanes, I. Mora-Jiménez, M. P. Martínez-Ruiz, J. L. Rojo-Álvarez, “Statistical nonlinear analysis for reliable promotion decision-making”. *Digital Signal Processing*, vol. 33, pp. 156-168, 2014.
- [5] C. Soguero-Ruiz, K. Hindberg, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R. O. Lindsetmo, K. M. Augestad, R. Jenssen. “Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records”. *IEEE Journal of Biomedical and Health Informatics*, PP (99) no. 1, 2014.

- [6] C. Soguero-Ruiz, A. Sánchez-Caro, I. Mora-Jiménez, P. Serrano-Balazote, L. Lechuga, J. Ramos-López, A. García-Alberola and J. L. Rojo-Álvarez. “Towards semantic interoperability in the electronic health record: a web prototype using cardiovascular archetypes”. (Submitted to PLOS ONE, January 2015).
- [7] C. Soguero-Ruiz, K. Hindberg, I. Mora-Jiménez, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R. O. Lindsetmo, K. M. Augestad, R. Jenssen. “Early detection of anastomosis leakage from heterogeneous data sources in electronic health records using kernel methods”. (Submitted to IEEE Journal of Biomedical and Health Informatics, January 2015).

International conferences

- [8] C. Soguero-Ruiz, F. J. Gimeno-Blanes, I. Mora-Jiménez, M. P. Martínez-Ruiz, J. L. Rojo-Álvarez. “Deal effect curve and promotional models-Using machine learning and bootstrap resampling test”. *The 1st International Conference on Pattern Recognition Applications and Methods*, vol. 2, pp. 537-540, 2012.
- [9] C. Soguero-Ruiz, L. Lechuga, I. Mora-Jiménez, J. Ramos-López, A. García-Alberola and J. L. Rojo-Álvarez. “Ontology for heart rate turbulence domain applying the conceptual model of SNOMED-CT”. *Computing in Cardiology*, vol. 39, pp. 89-92, 2012.
- [10] A. Sánchez-Caro, C. Soguero-Ruiz, I. Mora-Jiménez, L. Lechuga, J. Ramos-López, A. García-Alberola and J. L. Rojo-Álvarez. “Towards semantic interoperability for cardiovascular risk stratification into the electronic health records using archetypes and SNOMED-CT”. *Computing in Cardiology*, vol. 41, pp. 497-500, 2014.
- [11] C. Soguero-Ruiz, K. Hindberg, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R. O. Lindsetmo, K. M. Augestad, R. Jenssen. “Bootstrap resampling feature selection and support vector machine for early detection of anastomosis leakage”. *IEEE-EMBS International Conference on Biomedical and Health Informatics*, pp. 577-580, 2014. (Paper selected for extended version in the IEEE Journal of Biomedical Health Informatics).
- [12] C. Soguero-Ruiz, K. Hindberg, I. Mora-Jiménez, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R. O. Lindsetmo, K. M. Augestad, R. Jenssen. “Feature selection from bag-of-words in electronic health records for early detection of anastomosis leakage”. *2nd International Workshop on Pattern Recognition for Healthcare Analytics*, pp. 1-4, 2014.

- [13] C. Soguero-Ruiz, F. Wang, R. Jenssen, K. M. Augestad, J. L. Rojo-Álvarez, I. Mora-Jiménez, R. O. Lindsetmo, S. O. Skrøvseth. “Data-driven temporal prediction of surgical site infection”. AMIA 2015 (Submitted).
- [14] C. Soguero-Ruiz and et al. “Covariance kernel series”. (In preparation)

The research activity of this Thesis combines theoretical modeling and practical applications and has given rise to manuscripts with novelty in several fields. Thus, the contributions related to the general objective as well as with the specific objectives are next described.

- The contributions related to the *general objective* (presented in Chapter 2) of this Thesis are in [1], [2], [4], [5], [7], [11], [12], [14].
- The work developed within the *first specific objective* (presented in Chapter 3) has been published on [1], [2], [4], [8].
- The research developed within the *second specific objective* (presented in Chapter 4) has been published on [5], [7], [11], [12], [13], [14].
- The research developed within the *third specific objective* (presented in Chapter 5) has been published on [3], [6], [9], [10].

Theoretical Fundamentals and Contributions in Machine Learning

2.1 Introduction

The term *Machine Learning* (ML) has been widely studied in the literature for reproducing and improving the human capabilities to recognize patterns in the data by using automated and intelligent systems. Examples of applications using ML methods are marketing (e.g., sales promotion or client segmentation), web content search (e.g, social networks, page ranking or text categorization), or healthcare (e.g., diagnosis, early detection of complications, or phenotype discovery). ML methods allow to *learn* relationships among samples (observations, examples or data points) [4, 5, 6], each one described by a set of input features and the corresponding output. Towards that end, a statistical model is built to predict the desired output. If the output consists in one or more continuous variables, then the learning task is called regression [4]. When the output only consists in a finite number of discrete categories, it is called classification [4, 41].

In this Thesis, the ML predictive modeling schema presented in Fig 1.1 is followed [7], which consists of three different stages, namely, feature extraction, model design, and prediction model. The first stage, *feature extraction*, is very domain specific and often requires to be supported by domain experts. At this first stage, expert knowledge is needed to collect the features that are relevant to the problem and that can be used to feed the estimation model. Once the features have been extracted, the next stage is the *model design*, consisting of three different steps, namely, feature engineering, feature selection, and model exploration. Sometimes, data are sparse, high dimensional, scarce in terms of number of samples, or time dependent. In this scenario, a *feature engineering* stage is necessary to deal with missing values and to characterize the temporal dynamic of the features. Several strategies have been proposed ranging from simple

methods to sophisticated ones. *Feature selection* methods select a reduced number of features that maintain or improve the prediction model performance. In the *model exploration* step, a mathematical model is designed for prediction [4]. Finally, and using the built *prediction model*, expert interpretation and useful conclusions are obtained to support decision making in real-world applications.

In this Dissertation, several applications from widely different domains are addressed. However, the same ML techniques, and specifically the ones related to model design stage, can be used to analyze data from these different nature applications. Thus, the theoretical fundamentals and contributions in ML for feature engineering, feature selection, and model exploration, are first briefly described in this chapter. The complete predictive modeling pipeline is presented individually for each application in devoted chapters.

For the remaining of this Thesis, let $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$ denote the data set, where $\mathbf{x}_i \in \mathbb{R}^N$ and y_i is the observed output, being $y_i \in \{-1, +1\}$ for a binary classification task or a continuous value $y_i \in R$ for a regression task.

2.2 Feature Engineering

2.2.1 Background

Data are described by a number of binary, continuous, or categorical features. Sometimes these features can be sparse, high dimensional, scarce in terms of number of samples, and time dependent. In these cases, a *feature engineering* approach is required to characterize the dynamic of each feature. This stage is one of the major aspects to consider when building predictive models in domains with different nature.

Data preprocessing (feature engineering) is an important step and usually the most time consuming stage in the whole predictive modeling pipeline. The complexity of data preprocessing depends on the amount of redundant information and noise that are present in the data sources. Outlier removal, normalization, or missing values handling, are examples of data preprocessing. An outlier is an observation that is extremely distant from the other observations and that be due to variability in the measurement. As a rule of thumb, sometimes a threshold based on a number of times the standard deviation is used to roughly identify them [42], but this is not a statistically founded criterion, and rather the expert inspection is the approach to be followed in most cases to identify and deal with outliers. They can seriously distort the learning process, thus, outliers are normally removed. On the other hand, some ML methods are very sensitive to the chosen scale of input variables (e.g., distance-based methods) since the influence of each variable can be different. To avoid that, a normalization step is normally considered independently for each variable, for example, by transforming each feature so that its statistic is mean zero and standard

deviation one.

When using observational data from secondary sources such as the EHR, one also needs to take into account that data are usually sparse and irregularly sampled, i.e., certain features for some samples are missing. For example, blood tests are taken at a mixture of predefined stages in a patient pathway and by a clinically driven sampling. Thus, if predictive analytics relies on regularly sampled data, specific imputation methods need to be employed such that regular sampling is retrieved, and hence, sparse data can be treated as missing data. Most traditional techniques for dealing with missing data include replacing the missings with zero values, with the mean of the available values of each feature, with the last observed value (called Last Observation Carried Forward, LOCF), or based on the k nearest neighbors technique, as proposed in [43]. Alternatively, Lasko et al. [44] suggested using Gaussian Processes (GP) followed by a warped function as a methodology to deal with sparse data. The warped function is intended to adjust for the fact that rapid changes in temporal variables in connection with active treatment is often followed by long periods of apparent stability, leading to highly non-stationary processes. The time warping function can be constructed as

$$d' = d^{1/\alpha} + \beta \quad (2.1)$$

where d is the original distance between two adjacent observations, and α and β are free parameters to be tuned. This function converts non-stationary data into a stationary process which allows the use of a GP to deal with sparsity. GP are described in detail in Sec. 2.3.1.

Thus, it is a challenge working with data characterized by sparsity, irregular sampling, temporal structure and changing dynamics. New strategies are required, at least, to evaluate the performance and the information provided by each feature individually independently of the imputation method considered. The proposal of a temporal statistical analysis to individually characterize each feature is here addressed, aiming to provide with the following advantages in our different application scenarios: (1) more knowledge about the temporal dynamics of each feature; (2) a comparison among the behavior of the features and the previous studies in the literature; (3) the temporal trend of each feature, in some cases, before and after a reference time point (e.g., surgical intervention); and (4) a tool to define the uncertainty of each feature, specifically, after applying methods to deal with missing values when data are sparse and irregularly sampled.

The reconstruction of input spaces with time dependence requires to deal with FS and imputation methods, while maintaining the temporal properties. To deal with it, statistical moments and other summary parameters can be considered as inputs to the predictive models, however, temporal information is certainly lost. In this work, we pay special attention to the temporal properties of the input space to be conserved as much as possible, and two different approaches are considered. On the one hand, we characterize the temporal evolution of each feature with its mean and confidence interval (CI) after considering an imputation approach. On

the other hand, once a multivariable model has been built, it can be difficult to characterize the complex interactions and the underlying statistical properties among the inputs and the output. Thus, the second approach consists in studying the distribution of the output conditioned to the input features.

In this work, the use of nonparametric resampling techniques is proposed for providing with statistical processing methods to characterize the uncertainty, either in input spaces individually, or for them jointly with the output.

2.2.2 Contribution 1. Statistical Characterization of Features

In this Thesis, two approaches are proposed to define statistically each feature. In the first one, a statistical method is proposed for analyzing individually temporal features, whereas in the second one, the joint distribution of the input feature with the desired output is analyzed by considering a nonlinear multidimensional model, as described next.

Statistical Method for Characterizing Temporal Features

A nonparametric approach based on bootstrap resampling is proposed to individually characterize the temporal statistics of the j -th feature. Bootstrap resampling techniques can provide a useful framework for empirical and nonparametric estimation of the probability density function (*pdf*) of statistical entities from a set of observations [45].

Let $\mathbf{x}^{j(t)}$ ($j = 1, \dots, N; t = 1, \dots, T$) be a feature vector where $\mathbf{x}^j \in R^T$, with T the number of samples for a given temporal feature at a regularly sampled grid. Its statistical distribution is defined as $p_{\mathbf{x}^{j(t)}}$ and can be approximated by an empirical estimation, obtained from sampling with replacement the set of observations in $j(t)$. Thus, a new set $X^{*,j(t)}$ is first built, where superscript * represents, in general terms, any observation of the feature j in the time t from the bootstrap resampling process. Therefore, the set $X^{*,j(t)}$ contains elements of $\{\mathbf{x}_i^{j(t)}\}_{i=1}^n$ which are included none, one, or several times. The resampling process is repeated B times, yielding $\{X^{*,j(t)}(b)\}_{b=1}^B$. A *bootstrap replication* of an estimator is obtained by using a given operator with the elements in the bootstrap resample, so that the bootstrap replication of the statistical magnitudes of interest is given by $\hat{x}^{*,j(t)}(b) = F(\mathbf{x}^{*,j(t)}(b))$. Statistic operator $F(\cdot)$ can be used to estimate the statistical distribution of the replicated magnitude, such as the average and the standard error. Note that this procedure respects the possible presence of temporal dynamics in data.

Distribution of the Output using the Input Features

On the other hand, and again based on bootstrap technique, the statistical distribution of the output as a function of the input features can be obtained as $p_{y(\mathbf{x})}$. In our scenario, estimating

this multivariate distribution allows us to characterize multidimensional feature spaces by simply bootstrapping the available observations.

Let $V = \{\mathbf{x}_i, y_i\}_{i=1}^n$ be a set consisting of the input and output vectors. Thus, $p_{y(\mathbf{x})}$ is approximated by an empirical estimation, obtained from sampling with replacement observations in V . First, a new set $V^* = \{\mathbf{x}_i^*, y_i^*\}_{i=1}^n$ is built, where superscript $*$ represents, in general terms, any observation from the bootstrap resampling process. Therefore, set V^* contains elements of V which are included none, one, or several times. We repeat the resampling process B times, yielding $\{V^*(b)\}_{b=1}^B$. A bootstrap replication of an estimation u is calculated from the observations in the bootstrap resamples, thus is, $u^*(b) = F(V^*(b))$. Then, we can estimate $p_{y(\mathbf{x})}^*$ for the statistical distribution $p_{y(\mathbf{x})}$, and used it to readily estimate the quality and the reliability of the output.

The influence of the simulated regular sampling to characterize individually each feature by its mean and CI obtained after considering a nonparametric resampling approach, is computed in Application 4.1 (Sec. 5.2) and in Application 4.3 (Sec. 4.4). The characterization of the output conditioned to the input features using a nonlinear multidimensional model is computed in Application 3.2 (Sec. 3.3).

2.3 Predictive Modeling

In this Thesis, several predictive models are first studied, ranging from classical methods to more complex ones, such as artificial neural networks (ANN) or GP. Then, a smoothing regression method based on the properties of the covariance matrix, called *Covariance Kernel Series*, is proposed

2.3.1 Machine Learning Methods

There are in the ML literature several classification and regression methods for linear and nonlinear tasks. In this Thesis, Fisher's discriminant analysis (FDA), naive bayes (NB) and support vector machines (SVM) are briefly described for classification, whereas k -nearest neighbors (k -NN), general regression neural networks (GRNN), multilayer perceptron (MLP), SVM, and GP are studied for regression. Furthermore, multi source and composite kernels are presented.

The general linear estimation model is given by $y = \langle \mathbf{x}, \mathbf{w} \rangle + b$, where \mathbf{x} is the input (column) vector, \mathbf{w} is the weight vector, b is the bias term, y is the desired output and $\langle \cdot, \cdot \rangle$ denotes the inner product.

Fisher Criterion. The goal of FDA in the two-class problem [46] is to find a discriminating

linear projection $\langle \mathbf{w}, \mathbf{x} \rangle$, by simultaneously maximizing the between-class scatter and minimizing the within-class scatter on the projected output given by the cost function

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_B \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_W \mathbf{w}} \quad (2.2)$$

where \mathbf{S}_B and \mathbf{S}_W denote the between-class scatter matrix and the within-class scatter matrix in the original space, and \top denotes transposed vector. These are defined by $\mathbf{S}_B = \sum_{c=1}^2 n_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top$ and $\mathbf{S}_W = \sum_{c=1}^2 \sum_{i \in C_c} (\mathbf{x}_i - \boldsymbol{\mu}_c)(\mathbf{x}_i - \boldsymbol{\mu}_c)^\top$, respectively. Here, n_c is the number of samples in class C_c , with $c = 1, 2$. Furthermore, $\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{i \in C_c} \mathbf{x}_i$ and $\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$. In order to classify the projected points, a threshold has to be determined. There is no general rule for finding this threshold, but a common choice is the average between the class-conditional means.

Naive Bayes. The NB classifier [18] estimates the class-conditional probability density functions assuming conditionally independent features, i.e.,

$$p(\mathbf{x}|y = c) = \prod_{m=1}^N p(x^m|y = c) \quad (2.3)$$

where $\mathbf{x} = [x^1, \dots, x^N]^\top$ is the input feature vector and $c = 1, 2$ denotes the class.

The model is called *naive* since input features are expected to be independent, even conditional on the class label. Despite this assumption, classifiers based on NB have been successful in many applications, sometimes giving competitive results with respect to other more sophisticated methods [47, 48].

k -NN. The k -NN is a nonparametric procedure which provides an estimation of the output, $f(\mathbf{x}_*)$, from the k input samples in the training set closest to \mathbf{x}_* according to a measurement of similarity or distance [18, 49]. Conventional distance measurements are L1 and L2 norms, and many different measurements have been proposed according to the nature of the data [18]. The k -NN estimator output is given by

$$y = f(\mathbf{x}_*) = \frac{\sum_{i=1}^k w_i y_i}{\sum_{i=1}^k w_i} \quad (2.4)$$

where w_i is a weighting function that depends on the distance of new input sample \mathbf{x}_* to the i -th nearest training sample, and parameter k has to be previously fixed during the design procedure.

MLP based models. ANNs are multiparametric nonlinear models, capable of learning from samples and discovering complex relationships among variables. Neurons are the basic elements

of ANN, and they represent simple, highly interconnected processing units, usually grouped in several layers (input layer, hidden layers, and output layer). Each interconnection has a numeric weight, which has to be adjusted during the training stage. In the MLP network, one of the mostly used ANN, there is one input neuron for each variable in input sample \mathbf{x} , and as many output neurons as output variables to be estimated (i.e., \mathbf{y} can be a multivariable output). Hence, the number of hidden layers and the number of neurons in each have to be chosen during the design process. Hidden layer neurons in MLP correspond to global functions, so-called activation functions, such as linear or sigmoid, and the MLP is a universal approximator (a single hidden layer is capable of approximating any continuous, smooth, and bounded function) [50]. During the learning process, weights among neurons connections $\{\mathbf{w}\}$ are adjusted, according to a given cost function. The most widespread training algorithm is back-propagation [4], which consists of an iterative process starting in a given initial solution and a gradient-descent optimization based on first-order derivatives. For nonconvex functions, local minima can be present, which can be alleviated by the consideration of the second-order derivatives, as for instance in the Levenberg-Marquardt algorithm [51].

GRNN based models. Another ANN which has received much attention, also used in this work, is the GRNN, a nonparametric estimator given by the minimization of the squared error on the set of available examples [52]. Function $f(\mathbf{x})$ minimizing this error is

$$f(\mathbf{x}) = E[y|\mathbf{x}] = \frac{\int yp(\mathbf{x}, y)dy}{\int p(\mathbf{x}, y)dy} \quad (2.5)$$

where E denotes statistical expectation, and $p(\mathbf{x}, y)$ is the joint *pdf* of \mathbf{x} and y . Given that $p(\cdot)$ is often unknown, it can be estimated by using nonparametric estimation techniques, such as Parzen windows with Gaussian kernels. In this case, the GRNN estimator is given by

$$f(\mathbf{x}) = \frac{\sum_{i=1}^n y_i e^{-\frac{D_i^2}{2\sigma^2}}}{\sum_{i=1}^n e^{-\frac{D_i^2}{2\sigma^2}}} = \sum_{i=1}^n h_i y_i \quad (2.6)$$

where σ is the kernel width, $D_i^2(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_i)^\top (\mathbf{x} - \mathbf{x}_i)$ is the squared Euclidean distance between input sample \mathbf{x} and design example \mathbf{x}_i . For high values of the kernel width, the output depends on too distant examples, and it is an over-smoothed estimation of the actual value, whereas for very low values, the network limits to estimate the value from the closest example to \mathbf{x} [52]. Parameter σ has to be tuned during the training procedure.

SVM for classification and regression. We focus first on the SVM classifier (see e.g. [53, 54]), integrating regularization in the same classification procedure, such that model complexity is controlled, and the upper bound of the generalization error is minimized. These theoretical properties make the SVM an attractive approach for several data tasks.

The SVM classification algorithm seeks the separating hyperplane with the largest margin between two classes. The hyperplane optimally separating the data is defined from a subset of training data (also called support vectors), and it is obtained by minimizing $\|\mathbf{w}\|^2$, as well as the classification losses in terms of a set of slack variables $\{\xi_i\}_{i=1}^n$. Considering the ν -SVM introduced by Schölkopf et al. [55] and a potential nonlinear mapping $\phi(\cdot)$, the ν -SVM classifier solves

$$\min_{\mathbf{w}, \{\xi_i\}, b, \rho} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 + \nu\rho + \frac{1}{n} \sum_{i=1}^n \xi_i \right\} \quad (2.7)$$

subject to:

$$y_i(\mathbf{w}^\top \phi(\mathbf{x}_i) + b) \geq \rho - \xi_i \quad \forall i = 1, \dots, n \quad (2.8)$$

$$\rho \geq 0, \quad \xi_i \geq 0 \quad \forall i = 1, \dots, n \quad (2.9)$$

Variable ρ adds another degree of freedom to the margin, and the margin size linearly increases with ρ . Parameter $\nu \in (0, 1)$ acts as an upper bound on the fraction of margin errors, and it is also a lower bound on the fraction of support vectors. Appropriate choice of nonlinear mapping ϕ guarantees that the transformed samples (input vector) are more likely to be linearly separable in the (higher dimensional) feature space.

The primal problem in Eq. (2.7) is solved by using its dual formulation, yielding $\mathbf{w} = \sum_{i=1}^n y_i \alpha_i \phi(\mathbf{x}_i)$ (see [54] for further details), where α_i are Lagrange multipliers corresponding to constraints in Eq. (2.8). Thus, the decision function for any test vector \mathbf{x}_* is given by

$$f(\mathbf{x}_*) = \left(\sum_{i=1}^n y_i \alpha_i K(\mathbf{x}_i, \mathbf{x}_*) + b \right) \quad (2.10)$$

In order to predict the label of \mathbf{x}_* , the sign of $f(\mathbf{x}_*)$ is used. The so-called support vectors are those training samples \mathbf{x}_i with corresponding Lagrange multipliers $\alpha_i \neq 0$. The bias term b is calculated by using the *unbounded* Lagrange multipliers as $b = \frac{1}{k} \sum_{i=1}^k (y_i - \mathbf{w}^\top \phi(\mathbf{x}_i))$, where k is the number of non-null and unbounded Lagrange multipliers.

The use of Mercer kernels allows to handle the nonlinear algorithm implementations as $K(\mathbf{x}_i, \mathbf{x}_*) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_*) \rangle$. In this work, two well-known Mercer kernels are used: the linear kernel, given by $K(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle$, and the Radial Basis Function (RBF) kernel, given by $K(\mathbf{x}, \mathbf{z}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}\right)$, where σ is the width parameter, to be tuned together with free parameter ν .

Given a test sample \mathbf{x}_* , the traditional SVM classifies it according to the value of decision function $f(\mathbf{x}_*)$. However, it is also possible to convert the output of the classifier into a posterior probability of class membership by using a sigmoidal function mapping approach [56] as follows,

$$Pr(y = 1|\mathbf{x}_*) \approx \frac{1}{1 + \exp(af + c)} \quad (2.11)$$

where $f = f(\mathbf{x}_*)$, and a and c are estimated by minimizing the negative log-likelihood function (see [56] and references therein for details).

Conventional SVM regression uses the regularized ϵ -insensitive cost (or Vapnik's cost) [53]. Parameter ϵ has not a compact support, and then its practical tuning can become inaccurate, resulting in an extensive scanning for the cases with unknown accuracy of the approximation. Alternatively, the ν -SVM has been proposed for automatically tuning ϵ through a new free parameter ν with bounded range $(0, 1)$ [57]. The ν -SVM algorithm for non linear regression optimizes the following primal functional for ϵ -insensitive cost:

$$L = \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\nu \epsilon + \frac{1}{N} \sum_{i=1}^n (\xi_i + \xi_i^*) \right) \quad (2.12)$$

where ξ_i, ξ_i^* are the slack variables, C is the regularization parameter, and ν allows to give an approximate ratio of the number of support vectors with respect to the number of training examples. The following constrains must hold:

$$(\mathbf{w}^\top \mathbf{x}_i + b) - y_i \leq \epsilon + \xi_i \quad (2.13)$$

$$y_i - (\mathbf{w}^\top \mathbf{x}_i + b) \leq \epsilon + \xi_i^* \quad (2.14)$$

and $\xi_i, \xi_i^* \geq 0, \epsilon \geq 0$ for $\forall i = 1, \dots, n$. The Lagrangian functional can be written, by using Lagrange multipliers $\alpha, \alpha^*, \eta, \eta^*$ and β , given

$$\begin{aligned} L = & \frac{1}{2} \|\mathbf{w}\|^2 + C\nu\epsilon + \frac{C}{N} \sum_{i=1}^n (\xi_i + \xi_i^*) - \beta\epsilon - \sum_{i=1}^n (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^n \alpha_i (\xi_i + y_i - \mathbf{w}^\top \mathbf{x}_i - b + \epsilon) - \sum_{i=1}^n \alpha_i^* (\xi_i - y_i + \mathbf{w}^\top \mathbf{x}_i + b + \epsilon) \end{aligned} \quad (2.15)$$

By minimizing this functional with respect to primal variables, the Karush-Khun-Tucker conditions are obtained, and after their substitution, the final solution is given by

$$f(\mathbf{x}) = \left[\sum_{i=1}^n (\alpha_i^* - \alpha_i) \mathbf{x}_i^\top \right] \mathbf{x} + b \quad (2.16)$$

Dual variables α_i and α_i^* will be nonzero whenever samples \mathbf{x}_i give a residual either in the boundary or out of the insensitivity region. By introducing the nonlinear mapping and then substituting the dot products by kernel functions the following dual problem is obtained:

$$\max_{\alpha_i, \alpha_i^*} \sum_{i=1}^n (\alpha_i - \alpha_i^*) y_i - \frac{1}{2} \sum_{i,j=1}^n (\alpha_i^* - \alpha_i) (\alpha_j^* - \alpha_j) K(\mathbf{x}_i, \mathbf{x}_j) \quad (2.17)$$

constrained to

$$\sum_{i=1}^n (\alpha_i - \alpha_i^*) = 0, \quad \sum_{i=1}^n (\alpha_i - \alpha_i^*) \leq C\nu \quad (2.18)$$

$$\alpha_i, \alpha_i^* \in \left[0, \frac{C}{n} \right] \quad (2.19)$$

The nonlinear estimator has the following form:

$$f(\mathbf{x}) = \sum_{i=1}^n (\alpha_i^* - \alpha_i) K(\mathbf{x}, \mathbf{x}_i) + b \quad (2.20)$$

where b is obtained from Eq. (2.13) and (2.14) when $\xi_i = \xi_i^* = 0$.

From above summary of the ν -SVM algorithm for both regression and classification, it is clear that there are several free parameters to be tuned. With respect to the cost function, parameter ϵ can be readily substituted by ν , and C is the linear cost parameter which is only tuned in classification tasks. With respect to the kernels, parameter σ has to be previously tuned when using a RBF kernel.

Sources fusion using kernels. The performance of a prediction system can be improved by including heterogenous data sources. One influential way to do this is by exploiting the so-called composite kernels, which combine different kernels, each associated with a different data source. Some properties of Mercer's kernels are relevant for this work. Let K_1 and K_2 be Mercer kernels over $\mathcal{X} \times \mathcal{X}$, with $\mathbf{x}, \mathbf{z} \in \mathcal{X} \subseteq \mathbb{R}^N$. Then, the following are valid Mercer's kernels [58],

$$K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z}) + K_2(\mathbf{x}, \mathbf{z}) \quad (2.21)$$

$$K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z}) \cdot K_2(\mathbf{x}, \mathbf{z}) \quad (2.22)$$

$$K(\mathbf{x}, \mathbf{z}) = \mu K_1(\mathbf{x}, \mathbf{z}) \quad (2.23)$$

$$K(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{A} \mathbf{z} \quad (2.24)$$

where \mathbf{A} is a symmetric positive semi-definite ($N \times N$) matrix, and $\mu > 0$. The Mercer's kernels properties, together with simple vector concatenation, allow us to create composite kernels in several ways [58]. This gives a framework for exploring the most convenient way of combining different data sources. Among them, the stacked and composite kernels are next described.

Stacked Kernel. A common way to combine data is obtained by following a stacked approach. The main idea of the stacked input vector kernel [58] consists in merging different data sources \mathbf{x}_i^s , where $s = 1, \dots, S$, being S is the number of sources. The new input vector $\tilde{\mathbf{x}}_i$ is given by

$$\tilde{\mathbf{x}}_i = [(\mathbf{x}_i^1)^\top, (\mathbf{x}_i^2)^\top, \dots, (\mathbf{x}_i^S)^\top]^\top, \quad (2.25)$$

and its dimension is obtained as the sum of dimensions of the S sources under consideration. Then, a single kernel can be used, given by

$$K_{st} = K(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) \quad (2.26)$$

Composite Kernels. Commonly, input data are originated from sources of different nature. A viable approach is to affiliate different kernels to each source, and combine them using a

composite kernel approach [58, 59, 60]. A simple composite kernel combining heterogenous data sources can be obtained by concatenating linear and nonlinear transformation for each \mathbf{x}_i^s . Let $\varphi(\cdot)$ be a linear or a nonlinear transformation into its corresponding Hilbert space \mathcal{H} , and let \mathbf{A}_s be a linear transformation from \mathcal{H}_s to \mathcal{H} , respectively. Thus, the mapping to \mathcal{H} can be described as follows:

$$\phi(\mathbf{x}_i) = \{\mathbf{A}_1\varphi_1(\mathbf{x}_i^1), \mathbf{A}_2\varphi_2(\mathbf{x}_i^2), \dots, \mathbf{A}_S\varphi_S(\mathbf{x}_i^S)\} \quad (2.27)$$

and the corresponding inner product can be easily computed:

$$K_{ck}(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle = \sum_{s=1}^S \varphi_s(\mathbf{x}_i^s)^\top \mathbf{A}_s^\top \mathbf{A}_s \varphi_s(\mathbf{x}_j^s) = \sum_{s=1}^S K_s(\mathbf{x}_i^s, \mathbf{x}_j^s) \quad (2.28)$$

where the property from Eq. (2.24) is exploited in the last step. Previous composite kernel is a simple sum of the individual samples' kernel-based similarities for each data source, and is known to be robust against overfitting. Furthermore, in the *weighted* summation kernel, the importance of each data source can be modified by further exploiting Eq. (2.23), yielding

$$K_{ws}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^S \mu_s K_s(\mathbf{x}_i^s, \mathbf{x}_j^s) \quad (2.29)$$

where weight μ_s gives different relevance to each data source. In this work, each μ_s is a free parameter to be tuned.

Gaussian Process Regression. A random process $f(\mathbf{x})$ is a GP if, for any finite set of values of $\{x_1, x_2, \dots, x_n\}$, the variables of the corresponding random vector $\mathbf{f} = [f(x_1), f(x_2), \dots, f(x_n)]^\top$ are jointly normal (Gaussian). Element K_{ij} of the covariance matrix \mathbf{K} of \mathbf{f} is $k[f(x_i), f(x_j)]$ where $k[\cdot, \cdot]$ is a covariance (kernel) function, such as the RBF, or the squared exponential function. Using Bayes Theorem, the posterior density function for random variable $f_* = f(x_*)$ conditioned on the observed \mathbf{f} becomes

$$\mathbf{p}(f_* | \mathbf{f}) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp \left[-\frac{(f_* - \hat{f})^2}{2\hat{\sigma}^2} \right], \quad (2.30)$$

where the posterior mean value is given by $\hat{f} = \mathbf{k}^T \mathbf{K}^{-1} \mathbf{f}$; and the posterior variance is $\hat{\sigma}^2 = \kappa - \mathbf{k}^T \mathbf{K}^{-1} \mathbf{k}$. In this expression, element i of the vector \mathbf{k} is $k[f(x_*), f(x_i)]$, $\forall i = 1, \dots, n$, and $\kappa = k[f(x_*), f(x_*)]$. In GP regression, \hat{f} is used as the estimate, or prediction, of f_* , while $\hat{\sigma}^2$ provides the level of confidence in the prediction.

2.3.2 Contribution 2. Covariance Kernel Series for Regression

The aim of this contribution is to establish a theory for covariance smoothed weakly stationary stochastic processes over \mathbb{R}^N to be used in regression problems. While classical theory of

stochastic process deals with random variables over time, we will focus on a potentially broader class of stochastic processes, inspired by covariance properties used in GPs. Let $F(\mathbf{x})$ be a wide sense stationary real stochastic process where the stochastic process $F(\mathbf{x})$ at location \mathbf{x} represents a random variable. The feature vector $\mathbf{x} \in X$ is the set of possible inputs, which could be more general than time. The goal of this contribution consists in approximating F by $S(\mathbf{x}) = \int wF(\mathbf{x}')d\mathbf{x}'$, considering a weighting factor w defined as a smoothing correlation term, as follows

$$w = k_F(\mathbf{x}, \mathbf{x}') = E[F(\mathbf{x})F(\mathbf{x}')] = k_F(\mathbf{x} - \mathbf{x}'), \quad (2.31)$$

where k_F is the autocorrelation function. The covariance function must be symmetric positive semi-definite (psd) such that $k_F(\mathbf{0}) = \max_{\mathbf{x}} k_F(\mathbf{x})$.

In practice, we have available a set D of n observations $D = \{(\mathbf{x}_i, \mathbf{f}_i) | i = 1, \dots, n\}$, where \mathbf{x} denotes an input vector (covariates) of dimension N and where the realization $f = f(\mathbf{x})$ of $F(\mathbf{x})$ constitute the dependent variable. The focal point of this study will be a stochastic process defined as a weighted average, i.e. a smoothing over $F(\mathbf{x})$, where the weighting is defined in terms of the covariance structure of $F(\mathbf{x})$ as follows,

$$S(\mathbf{x}) = \int k_F(\mathbf{x} - \mathbf{x}')F(\mathbf{x}')d\mathbf{x}', \quad (2.32)$$

There are several reasons why it may be interesting to study $S(\mathbf{x})$. First of all, if the underlying process $F(\mathbf{x})$ is corrupted by additive noise, i.e. $F(\mathbf{x}) + N(\mathbf{x})$, where $N(\mathbf{x})$ is a noise process, then the covariance smoothing may lower the influence of the noise. Secondly, when working with sparse data, covariance smoothing can be considered as an interpolation approach.

Mean and covariance properties of $S(\mathbf{x})$. Some properties of $S(\mathbf{x})$ are established below.

Mean Function. The mean function of $S(\mathbf{x})$ is given by

$$m_S(\mathbf{x}) = E \left[\int k_F(\mathbf{x} - \mathbf{x}')F(\mathbf{x}')d\mathbf{x}' \right] = \int k_F(\mathbf{x} - \mathbf{x}')E[F(\mathbf{x}')]d\mathbf{x}' = \int k_F(\mathbf{x} - \mathbf{x}')m_F(\mathbf{x}')d\mathbf{x}' \quad (2.33)$$

and $m_S(\mathbf{x}) = 0$ if $m_F(\mathbf{x}) = 0$ as assumed here.

Covariance function. The covariance function of $S(\mathbf{x})$ is given by

$$\begin{aligned} k_S(\mathbf{x}, \mathbf{x}') &= E \left[\int k_F(\mathbf{x} - \tilde{\mathbf{x}})F(\tilde{\mathbf{x}})d\tilde{\mathbf{x}} \int k_F(\mathbf{x}' - \check{\mathbf{x}})F(\check{\mathbf{x}})d\check{\mathbf{x}} \right] \\ &= \int \int k_F(\mathbf{x} - \tilde{\mathbf{x}})E[F(\tilde{\mathbf{x}})F(\check{\mathbf{x}})]k_F(\mathbf{x}' - \check{\mathbf{x}})d\tilde{\mathbf{x}}d\check{\mathbf{x}} \\ &= \int \int k_F(\mathbf{x} - \tilde{\mathbf{x}})k_F(\tilde{\mathbf{x}} - \check{\mathbf{x}})k_F(\mathbf{x}' - \check{\mathbf{x}})d\tilde{\mathbf{x}}d\check{\mathbf{x}} = k_S(\mathbf{x} - \mathbf{x}'). \end{aligned} \quad (2.34)$$

Eigenvalue and eigenfunction expansion of $S(\mathbf{x})$. Since the covariance function is psd, it may be expressed in terms of the following series expansion [61]:

$$k_F(\mathbf{x} - \mathbf{x}') = \sum_{k=1}^{N_F} \lambda_k \phi_k(\mathbf{x})\phi_k(\mathbf{x}'), \quad (2.35)$$

where N_F is the number of eigenvalues λ_k and eigenfunctions ϕ_k satisfying $\int k_F(\mathbf{x}-\mathbf{x}')\phi_k(\mathbf{x}')d\mathbf{x} = \lambda_k\phi_k(\mathbf{x})$. Hence, we have

$$S(\mathbf{x}) = \int \sum_{k=1}^{N_F} \lambda_k \phi_k(\mathbf{x}) \phi_k(\mathbf{x}') F(\mathbf{x}') d\mathbf{x}' = \sum_{k=1}^{N_F} \int \sqrt{\lambda_k} \phi_k(\mathbf{x}') F(\mathbf{x}') d\mathbf{x}' \sqrt{\lambda_k} \phi_k(\mathbf{x}) = \sum_{k=1}^{N_F} \beta_k \psi_k(\mathbf{x}) \quad (2.36)$$

where $\beta_k = \int \lambda_k \phi_k(\mathbf{x}') F(\mathbf{x}') d\mathbf{x}' = \sqrt{\lambda_k} \langle \phi_k(\mathbf{x}') F(\mathbf{x}') \rangle$ and $\psi_k(\mathbf{x}) = \sqrt{\lambda_k} \phi_k(\mathbf{x})$. Note that the process $F(\mathbf{x})$ enters explicitly only in the computation of the coefficient β_k . These coefficients weight the functions $\psi_k(\mathbf{x})$, depending on the covariates \mathbf{x} and the covariance function $k(\cdot, \cdot)$, constituting the orthogonal series in Eq. (2.36).

Analyzing function $\psi_i(\mathbf{x})$. A smoothing version of the process $F(\mathbf{x})$ is obtained by weighting it by the covariance function. In Eq. (2.35), the covariance function is expressed in terms of eigenvalues and eigenvectors. However, it is relevant to study what happens when the original process $F(\mathbf{x})$ is close to one of the eigenfunctions, i.e., $F(\mathbf{x}) \simeq \phi_k(\mathbf{x})$. Let $S(\mathbf{x})$ be defined as $S(\mathbf{x}) = \int k_F(\mathbf{x}-\mathbf{x}') F(\mathbf{x}') d\mathbf{x}'$, and following Eq. (2.36), the smoothing process can be expressed as:

$$S(\mathbf{x}) = \sum_{k=1}^{N_F} \sqrt{\lambda_k} \langle \phi_k(\mathbf{x}'), F(\mathbf{x}') \rangle \sqrt{\lambda_k} \phi_k(\mathbf{x}). \quad (2.37)$$

By definition: $\int \psi_j(\mathbf{x}) \psi_k(\mathbf{x}) d\mathbf{x} = 0$ and $\int \|\psi_k(\mathbf{x})\|^2 d\mathbf{x} = 1$, and considering them in Eq. (2.37) when $F(\mathbf{x}) = \phi_k(\mathbf{x})$, it is obtained,

$$S(\mathbf{x}) = \sum_{k=1}^{N_F} \sqrt{\lambda_k} \langle \phi_k(\mathbf{x}'), \phi_k(\mathbf{x}') \rangle \sqrt{\lambda_k} \phi_k(\mathbf{x}) = \sum_{k=1}^{N_F} \lambda_k \langle \phi_k(\mathbf{x}'), \phi_k(\mathbf{x}') \rangle \phi_k(\mathbf{x}) \quad (2.38)$$

Thus, if the eigenfunction represents the process itself, i.e., $F(\mathbf{x})$ is very close to $\phi_k(\mathbf{x})$ a scaling factor of $\frac{1}{\lambda_k}$ is necessary to obtain the original process.

Several conclusions can be obtained from these assumptions: (1) for $\beta_k = \lambda_k \langle \phi_k(\mathbf{x}'), F(\mathbf{x}') \rangle$, if β_k is small, it means that the eigenfunction is not able to represent the original process. Thus, the k -th component should not be considered to obtain a smoothed version of the original process $F(\mathbf{x})$; and (2) the scaling factor is only reasonable for high values of β_k , i.e., for eigenvalues which can represent the process; otherwise, only noise will be add to the system.

Empirical estimation from data. Given covariates \mathbf{x}_i for $i = 1, \dots, n$, λ_i and $\phi_i(\mathbf{x})$ may be estimated from the eigenvalues δ_i and eigenvectors \mathbf{e}_i of the psd covariance (kernel) matrix $\mathbf{K} : \mathbf{K}_{ij} = k(\mathbf{x}_i - \mathbf{x}_j)$, $i, j = 1, \dots, n$; yielding $\lambda_i \approx \frac{\delta_i}{n}$ and $\phi_i(\mathbf{x}) \approx \sqrt{n} e_{i,t}$, where $e_{i,t}$ is the t -th element. Hence, the estimation of $S(\mathbf{x})$ is given by

$$\hat{S}(\mathbf{x}) = \sum_{k=1}^{N_F} \hat{\beta}_k \sqrt{\delta_k} e_{k,t}, \quad (2.39)$$

where $\hat{\beta}_k = \sqrt{\delta_k} \mathbf{f}^\top \mathbf{e}_k$ and $\mathbf{f} = [f_1, \dots, f_n]^\top$.

Out of sample extension. Using the covariance smoothing approach proposed in this work, the estimate for a test input \mathbf{x}_* , when the labels for the training examples are known, is computed. Although mapped samples $\phi(\mathbf{x})$ are unknown, the projection $\pi_k(\mathbf{x}^*)$ of a testing point \mathbf{x}_* can be obtained as the inner product of $\phi(\mathbf{x}_*)$ with the eigenvector \mathbf{u}_k of the covariance matrix. Hence, following Nystrom approximation [62, 63, 64]:

$$\hat{\phi}_i(\mathbf{x}_*) = \frac{\sqrt{n}}{\delta_k} \sum_{i=1}^n e_{i,t} K(\mathbf{x}_i, \mathbf{x}_*), \quad (2.40)$$

Then, the estimate for a test input \mathbf{x}_* is given by,

$$\begin{aligned} \hat{S}(\mathbf{x}_*) &= \sum_{k=1}^{N_F} \sqrt{\delta_k} \mathbf{f}^\top \mathbf{e}_k \sqrt{\lambda} \phi_k(\mathbf{x}_*) \\ &= \sum_{k=1}^{N_F} \sqrt{\delta_k} \mathbf{f}^\top \mathbf{e}_k \sqrt{\frac{\delta_k}{n}} \frac{\sqrt{n}}{\delta_k} \sum_{i=1}^n e_{i,t} K(\mathbf{x}_i, \mathbf{x}_*) \\ &= \sum_{k=1}^{N_F} \mathbf{f}^\top \mathbf{e}_k \sum_{i=1}^n e_{i,t} K(\mathbf{x}_i, \mathbf{x}_*) \end{aligned} \quad (2.41)$$

Since the covariance function is psd, it can be estimated by the covariance (kernel) matrix as:

$$k_F(\mathbf{x} - \mathbf{x}') = K(\mathbf{x}, \mathbf{x}'). \quad (2.42)$$

The kernel matrix is commonly computed based on a parametrized function such as RBF, being this one used in this Thesis.

Experiments. The proposed method has been evaluated in two databases previously analyzed by GP. In the first one, we study a seven-degrees-of-freedom SARCOS anthropomorphic robot arm (downloaded from <http://www.gaussianprocess.org/gpml/>). This data set was used for regression tasks in [65, 66]. We have in this case $D = \{(\mathbf{x}_i, f_i) | i = 1, \dots, 1000\}$ where each \mathbf{x} input vector is 21-dimensional (7 joint positions, 7 joint velocities, 7 joint accelerations), and the target variable $f = f(\mathbf{x})$ is one of the 7 joint torque. The main idea here is to illustrate that information related to the process $F(\mathbf{x})$ may be extracted from the process $S(\mathbf{x})$, depending solely on the covariate \mathbf{x} and covariance function $k(\cdot, \cdot)$. In the second analyzed database, the training data were collected during the *SPARC* – 2003 and *SPARC* – 2004 campaigns, in Barrax, La Mancha in Spain [67]. The output training data is the actual chlorophyll content. The chlorophyll content was measured for certain crops (garlic, alfalfa, onion, sun power, corn, potato, sugar beet, vineyard and wheat) in Barrax. In both examples, Covariance Kernel Series method performs better in terms of accuracy than GP. Covariance Kernel Series method was used in Application 3.1 (Sec. 3.2), but the performance was worse than the one obtained with SVM. The scarcity in the number of samples and the dimensional of the data, having only 43

examples and 8 features, may provide these results. Future work may improve the previous limitations.

2.4 Feature Selection

2.4.1 Conventional Methods

Feature selection is defined as a series of actions in order to choose a subset of features that are relevant, while holding or improving the learning method performance. The task of FS is well known in the ML literature (for a review, see [6, 68, 69]) and it is specially relevant when working with high dimensional input spaces, due to: (1) the computational complexity, when the number of features is larger, the number of parameters (as weights in the linear SVM or neurons in the MLP) is also larger, and thus, the time and the complexity for designing the estimation model also increase; (2) the generalization problems, the higher the number of training samples related to the number of free parameters in the estimation model, the lower the possibility of overfitting the model; (3) mutual correlation, one feature can add value to the predictive model when it is analyzed individually, however, the information carried by this feature can be lower in combination with another one.

Three different types of FS are common in the literature [69]. First, filter methods select features as a pre-processing step performed independently of the classifier. Second, wrapper methods evaluate the performance of the classifier based on subsets of features. An third, embedded methods integrate FS and classifier performance into the training procedure of the classifier [68, 69]. Examples of previous FS methods range from feature-ranking techniques based on correlation, to sensitivity analysis [70], and to maximum margin criteria [68, 71]. FS in text documents have focused on criteria such as the document frequency, the term frequency, mutual information, information gain, odds ratio, χ^2 statistic, and term strength, to name a few [72, 73, 74].

FS set depends on both the method used to select the relevant features, and on the selection criterion to select them. In the FS literature, some works considered a criterion which attempt to maximize the class separability [68], whereas in others, the criterion tried to retain the discriminating power of the data defined by original features [75]. Thus, random subsets can be obtained depending on both the method and the criterion considered. To our best knowledge, there are no studies which try to defined this randomness. Thus, in this work, three different FS methods are proposed to deal with the randomness in this stage by taking advance of the statistical properties of the data.

Of particular interest is FS based on the weights obtained by a maximum margin SVM linear classifier, which we pursue in this exposition. There are several reasons for this: (i) the

robustness of the linear SVM in high-dimensional and noisy low sample size problems; and (ii) the one-to-one relationship between the weights of the linear classifier and the features (words), which enables the interpretation of the features. The latter is a significant advantage when compared to classifiers such as Gaussian maximum likelihood or ANN [76, 77], where the direct connections to the features are lost. The previous literature on SVM-based FS is to a large degree concentrated on the Recursive Feature Elimination (RFE) method [68], which has been shown to compare very favorably to many of the classical FS methods. RFE puts a threshold on the amplitudes of the weights obtained by the SVM. Hence, the user must either pre-specify the number of features to obtain, or alternatively, to engage in a computationally demanding cross-validation procedure, whereby features are eliminated recursively, thus requiring numerous SVM re-training procedures on subsets of features of decreasing size. This may be very time consuming, even for small sample sizes.

2.4.2 Contribution 3. Statistical Feature Selection Strategies

We propose a further research on FS strategies based on the statistical nature of the weights of the linear SVM, by investigating: (a) a simple statistical criterion based on leave-one-out; (b) an intensive-computation statistical criterion based on bootstrap resampling; and (c) an advanced statistical criterion based on kernel entropy, as explained below.

Leave One-Out Based Test

The Leave one-out (LOO) cross-validation method has been shown to give an almost unbiased estimator of the generalization properties of statistical learning models [78]. The concept can be used for estimating the *pdf* for each feature m .

The process is to create a matrix of weights \mathbf{W} with n rows and N columns, where n is the number of samples and N is the number of features. Each row of \mathbf{W} is a weight vector corresponding to the linear SVM solution by using LOO cross-validation. The LOO technique partitions the original data set into n subsets, one for validation and the remaining $n - 1$ for training. This process is repeated n times, setting apart for evaluation each of the n subsets just once, hence yielding \mathbf{W} . For the m -th feature with $m = 1, \dots, N$, a given linear classifier yields a weight vector \mathbf{w}^m , whose statistical distribution can be approximated with different empirical resampling criteria, denoted as $\hat{p}_{\mathbf{w}^m}$.

The estimated Confidence Interval (CI^m) is built for each \mathbf{w}^m , which has all the LOO estimations for the m -th feature, in order to determine whether this feature is relevant. Then, CI^m is used to perform a hypothesis test on the m -th feature, with $H_0 : 0 \in CI^m$ (feature m is irrelevant for the model) vs alternative hypothesis $H_1 : 0 \notin CI^m$ (feature m is relevant for the model).

Bootstrap Resampling-Based Test

Bootstrap resampling methods [79] are very useful approaches for nonparametric estimation of the distribution of statistical magnitudes. We propose a bootstrap resampling scheme (see Fig. 2.1) for building a statistical test for FS, as follows. We use \mathbf{W} to provide a statistical description of the noise assuming its variance is globally dependent on the weight magnitude, and locally constant for weights with similar magnitude. Under these conditions, for each feature m with $m = 1, \dots, N$, a local window of δ radius encompassing the 2δ nearest features is considered to build the set of weights given by $R^m = \{w_i^{m-\delta}, \dots, w_i^{m-1}, w_i^{m+1}, \dots, w_i^{m+\delta}\}_{i=1}^n$. Hence, R^m represents a noisy set of weights, with low (still non-null) probability of including representative weights. Then, for each m -th feature, the set $S^m = \{w_i^m\}_{i=1}^n$ represents the weight set to be tested for significance.

Weight sets R^m and S^m are used to estimate the marginal distribution of noisy and potentially relevant weights for the m -th input feature, respectively, by constructing bootstrap resamples. A *bootstrap resample* is a new set obtained from sampling with replacement the elements of the original set (R^m and S^m in our case), providing resamples $R^{*,m}$ and $S^{*,m}$, respectively. The resampling process is repeated B times, with b indexing the resampling number ($b = 1, \dots, B$). Thus, the b -th resamples $S^{*,m}(b)$ and $R^{*,m}(b)$ contain $2\delta n$ and n elements of S^m and R^m , respectively, appearing zero, one, or several times. A *bootstrap replication* of an estimator is constrained to the elements in the bootstrap resample. The bootstrap replication of the statistics of interest is $\Delta^{*,m}(b) = s^{*,m}(b) - r^{*,m}(b)$, where $s^{*,m}(b)$ and $r^{*,m}(b)$ are elements, randomly chosen, from $S^{*,m}(b)$ and $R^{*,m}(b)$, respectively. The B bootstrap replications for each feature m allow us to estimate the Confidence Interval ($CI^{*,m}$) for the statistics $\Delta^{*,m}$. Then, $CI^{*,m}$ is used to perform a hypothesis test on the m -th feature, with $H_0 : 0 \in CI^{*,m}$ (feature m is irrelevant for the model) vs alternative hypothesis $H_1 : 0 \notin CI^{*,m}$ (feature m is relevant for the model). Note that we only sample one pair of $s^{*,m}(b)$ and $r^{*,m}(b)$ for each b , producing one $\Delta^{*,m}(b)$ for each b , and that the process results in a feature being found relevant if it has a large absolute value compared to the mostly noise weights that have mostly smaller absolute weights.

Kernel Entropy Inference Test

The basic idea behind the proposed kernel entropy inference test for feature selection, is to select those features that correspond to the high entropy part of a *pdf*, describing a random variable considered to generate the features. The high entropy part of a *pdf* represents the most informative part, and it is associated with the tails of the *pdf*. Fig. 2.2 (a) illustrates a *pdf*, where the sum of the areas represented by the black regions represent the tail probability.

In order to achieve the entropy-based feature selection, we concentrate on Renyi's second

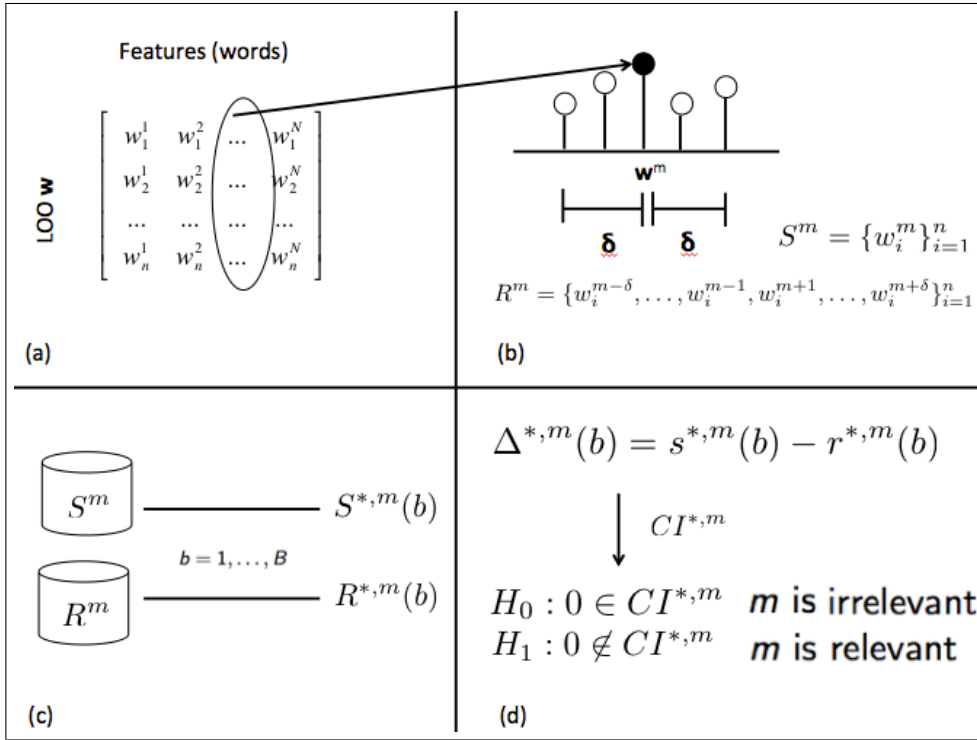


Figure 2.1: Schema of the proposed bootstrap resampling-based test: (a) Matrix of weights \mathbf{W} ; (b) 2δ nearest features represents a noisy set of weights R^m , whereas S^m is the weight set to be tested for significance; (c) bootstrap resamples; and (d) bootstrap replication and hypothesis test.

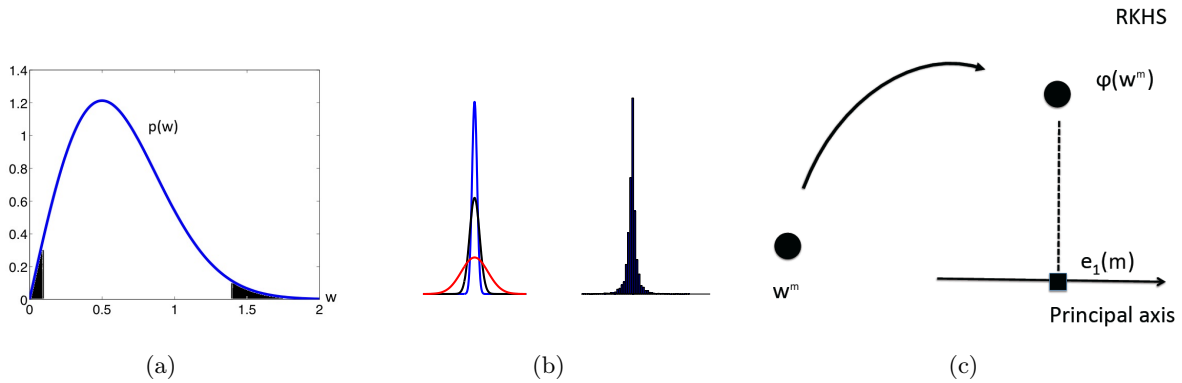


Figure 2.2: Kernel Entropy Inference Test: (a) The tail probability refers to the sum of the areas corresponding to the black regions under the probability density function $p(w)$. (b) Illustration of the role of the bandwidth, σ , in kernel density estimation (KDE). A large bandwidth (red) provides more smoothing compared to a small bandwidth (blue). (c) KECA is related to principal components in a RKHS corresponding to the positive semi-definite kernel function used in KDE.

order entropy [80] for a random variable w , given by

$$H(p) = -\log V(p), \quad V(p) = \int p^2(w')dw' \quad (2.43)$$

where $p(w)$ is the *pdf* of w . The reason for this choice is that this measure is easily estimated using the modern technique known as kernel entropy component analysis (KECA) [81]. KECA estimates the entropy using a kernel density estimator (KDE),

$$\hat{p}(w) = \frac{1}{N} \sum_{m=1}^N k_{\sigma}(w, w^m) \quad (2.44)$$

Here, w^m , $m = 1, \dots, N$, are elements of \mathbf{w} and the kernel function provides a smoothing of the histogram, where the bandwidth parameter σ governs the amount of smoothing. A common choice of kernel, which we also pursue in this paper, is $k_{\sigma}(w, w^m) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(w-w^m)^2}$. Figure 2.2 (b) illustrates the role of σ . A relatively big σ will tend to produce a too smooth density estimate and vice versa. Note that w is in this approach considered a one-dimensional random variable, and in that case reliable data-driven (automated) procedures exist for the selection of σ , meaning that a different σ is computed for different samples (data sets), see Section 4.2.3 for details. Furthermore, in the current exposition, the elements in the SVM weight vector \mathbf{w} represent the samples w^m of the random variable w . Based on one particular such \mathbf{w} , the left panel in Fig. 2.2 (b) (best viewed in color) shows the KDE based on an automated bandwidth selection procedure (blue), corresponding to the most narrow function shape. The broadest function (red) shows a Gaussian best fit. The right panel shows the histogram for \mathbf{w} indicating that the KDE performs better than the Gaussian model. In addition, the middle function (black) shows a kernel density estimate where we have manually doubled the selected σ . Note how the function becomes more smooth, in this case deviating more from the peaky shape.

When inserting Eq. (2.44) into Eq. (2.43), the KECA estimator for the Renyi entropy becomes $\hat{V}(p) = \frac{1}{N^2} \sum_{m=1}^N [\sqrt{\lambda_m} \mathbf{e}_m^{\top} \mathbf{1}]^2$. Here, λ_m and \mathbf{e}_m are eigenvalues and eigenvectors of the so-called kernel matrix \mathbf{K} where $K_{t,m} = k_{\sigma}(w^t, w^m)$ and $\mathbf{1}$ is a vector of ones. We have experienced robust estimates of $V(p)$ using only the top component (eigenvalue), such that in our case $\hat{V}(p) = [\sqrt{\lambda_1} \mathbf{e}_1^{\top} \mathbf{1}]^2$ (leaving out eigenvectors may be considered a de-noising process).

There is a one-to-one relationship between the elements in the vector \mathbf{e}_1 and the features stored in the SVM vector \mathbf{w} , and we use this in the FS. Since the kernel function is positive semidefinite, it computes an inner-product in a reproducing kernel Hilbert space (RKHS) [61]. That is, $w \mapsto \phi(w)$ such that the RKHS inner-product is $k_{\sigma}(w^t, w^m) = \langle \phi(w^t), \phi(w^m) \rangle$. It is furthermore known, that in RKHS, the projection of the j th point $\phi(w^m)$ equals $\mathbf{e}_1(m)$, i.e. the j th element of the eigenvector \mathbf{e}_1 . This is the RKHS principal component corresponding to $\phi(w^m)$. Hence, the feature w^m corresponds to the m -th element of \mathbf{e}_1 . This is illustrated in Fig. 2.2 (c).

The kernel entropy FS idea is the following. The tails of $p(w)$ contribute the most to the entropy of the random variable w and the features corresponding to the tail are represented by the smallest principal components in the RKHS (i.e, the smallest principal components contribute

the most to $\hat{V}(p)$). In the FS, we fix a tail probability, for example to the value 0.05, and select those features that correspond to the tail by identifying the corresponding smallest principal components (elements of \mathbf{e}_1). Note that the number of selected features by this proposed procedure is not pre-specified, but it depends on the chosen tail probability.

The FS strategies previously described are used in Application 4.1 (Sec. 5.2) and Application 4.2 (Sec. 4.3), see them for more details.

2.5 Model Selection

In this section, we first present several merit figures to evaluate the obtained model quality for both classification and regression methods. Next, the generalization capabilities of the developed models are analyzed. Finally, when more than one predictive model are evaluated, a statistical comparison is required to select the one which provides better performance. However, often, a description of the differences between them is not considered, or models are benchmarked using, t-student or ANOVA test or even nonparametric statistical tests such as Wilcoxon signed rank test [82]. In some studies, statistical assumptions of independence and gaussianity are not verified for its proper application. Therefore, this motivates the proposal of an operative benchmark methodology based on a cut-off nonparametric statistical test, both to characterize the generalization of the model as well as its comparison with other predictive models.

In this work, a nonparametric resampling test based on bootstrap is presented as a way to evaluate the models in terms of average and scatter measurements, for a more complete efficiency characterization of the predictive models. These statistical characterizations allow us to readily work with the distribution of the actual risk, in order to avoid overoptimistic performance evaluation in the ML based models. Apart from that, we propose a simple nonparametric statistical tool, based on the paired bootstrap resampling, to allow an operative result comparison among different learning-from-samples models. The use of bootstrap resampling in this setting is supported by the previous observation of heavy tails in the residuals distribution when using ML models, as well as by bimodalities, and other non-Gaussian effects [45], which make the use of conventional statistics a non-operative tool when working with ML models.

2.5.1 Merit Figures and Generalization Evaluation

Model quality obtained when applying learning techniques can be evaluated by means of informative merit figures. It is a well-known fact that the evaluation of merit figures in the training set is highly suboptimal, as far as generalization capabilities of the model are not considered at all. This is the main reason why it is necessary the adequate characterization of any merit figure for model benchmarking, and this characterization needs to be performed

using an independent data set of examples, i.e., a set n_{test} of data that not used during the training stage. Several merit figures can be used for benchmarking models in learning from samples techniques. We limit ourselves here to the Mean Absolute Error (MAE), given by

$$MAE = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} |f(\mathbf{x}_i) - y_i| \quad (2.45)$$

Performance measures in binary classification problem may be constructed based on the confusion matrix (CM), as follows,

		<i>Real diagnosis</i>	
		Positive	Negative
<i>Predictive diagnosis</i>	Positive	TP	FP
	Negative	FN	TN

where TP and TN denote true positives and true negatives, and FP and FN denote false positives and false negatives, respectively [83]. The performance measures considered in this work are the following: error probability, $P_e = \frac{FP+FN}{TP+FN+FP+TN}$; sensitivity, $Se = \frac{TP}{TP+FN}$; specificity, $Sp = \frac{TN}{TN+FP}$; and balanced error rate, $BER = \frac{1}{2}(Se + Sp)$.

Once the merit figure is selected, we need to analyze the generalization capabilities of the developed models. In that sense, cross-validation techniques are statistical methods for quantifying and measuring the generalization error [84]. In this work, a cross-validation technique is used for benchmarking ML techniques, by considering two subsets of the available examples, namely, a training set (for weights adjustment), and a validation set (for generalization benchmarking) [85]. Three widely used cross-validation techniques are: (a) Holdout, where data are split into 2 subsets for training (often 70%) and validation; (b) K-fold, where examples set is randomly divided into K subsets with the same size, one for validation and the remaining $K - 1$ for training, the process is repeated K times (each corresponding to a different subset being used for validation), and the generalization is obtained by averaging the merit figures of the K models; and (c) leave-one-out (LOO), a particular case of K -fold where K is the number of available examples. Note that the computational burden of LOO is much higher than other cross-validation techniques, however, in this work, we use LOO for the free parameter tuning in each technique, due to its advantage when the available data set is scarce.

2.5.2 Contribution 4. Bootstrap Resampling for Benchmarking Machine Learning Models

Actual Risk Estimation with Bootstrap Resampling

One of the main limitations of current ML techniques is the difficulty in establishing clear cut-off tests for model comparison, hence, systematic procedures for establishing FS, significance levels, and confidence intervals for model diagnosis, are developed. An interesting approach to the model diagnosis and FS can be given by bootstrap resampling techniques, which were first proposed as nonparametric procedures for estimating the *pdf* of a statistical estimation from a limited, yet informative enough, set of observations [45]. Bootstrap resampling has been successfully used before for selecting the design parameters of SVM classifiers [86], and due to their simplicity of use, we propose here to extend their use to model benchmarking in predictive modeling problems.

For a given set V of n observations, the dependence between the explanatory variables and the response variable can be fully described by means of the distribution of the output,

$$p_{y(\mathbf{x})} \rightarrow V = \{(y_i, \mathbf{x}_i); i = 1, \dots, n\}, \quad (2.46)$$

In order to obtain a ML model, a set of R weights $\{\omega_r, r = 1, \dots, R\}$, has to be estimated according to an optimization process denoted by operator $s(\cdot)$, and it depends on observations V and on the model design parameters that have been fixed a priori, which can be grouped in a vector $\boldsymbol{\theta}$ for a given ML technique. The model weights obtained by using the observations and a previously fixed $\boldsymbol{\theta}$ are given by

$$\omega = \{\omega_r\} = s(V, \boldsymbol{\theta}) \quad (2.47)$$

The model performance can be evaluated with the *empirical risk*, defined as a certain figure of merit of the model that is evaluated at the observations used for building the model, and it can be expressed as

$$\hat{R}_{emp} = t(\omega, V), \quad (2.48)$$

where $t(\cdot)$ represents the operator that stands for the figure of merit calculation.

Given that the ML based models do not rely on any a priori distribution of the data, it is not easy to know the functional form of the *pdf* of the merit figures. Moreover, the sample estimators of the merit figures can be optimistically biased, especially for some degenerate choices of the design parameters, e.g., when too much emphasis is put on the cost of the residuals, or when a too small neighborhood parameter is used. A method for estimating the *pdf* of the output is given by bootstrap resampling, and it can be used for compensating the optimistic bias in the figures of merit estimators [86].

A *bootstrap resample* is a data subset that is drawn from the observation set according to their empirical distribution $\hat{p}_{y(\mathbf{x})}$. Hence, the true *pdf* is approximated by the empirical *pdf* of the observations, and the bootstrap resample can be seen as a sampling with replacement process of the observed data, this is,

$$\hat{p}_{y(\mathbf{x})} \rightarrow V^* = \{(y_i^*, \mathbf{x}_i^*); i = 1, \dots, n\} \quad (2.49)$$

where superscript $*$ represents, in general terms, any observation, functional, or estimator arising from the bootstrap resampling process. Therefore, resampled set V^* contains elements of V which are included none, one, or several times. The resampling process is repeated B times. Accordingly, a partition of V in terms of resample V^* can be done, which is given by

$$V = \{V_{in}^*(b), V_{out}^*(b)\}_{b=1}^B, \quad (2.50)$$

where $V_{in}^*(b)$ and $V_{out}^*(b)$ are the subsets of observations that are and are not included in resample b , respectively.

A *bootstrap replication* of an estimator is given by its calculation constrained to the observations in the bootstrap resample. The bootstrap replication of the empirical risk estimator is

$$\hat{R}_{emp}^*(b) = t(\omega, V_{in}^*(b)). \quad (2.51)$$

The normalized histogram obtained from B resamples is an approximation to the *pdf* of the empirical risk. However, further advantage can be obtained by calculating the bootstrap replication of the risk estimator on the non-included observations, and rather than estimating the empirical risk, we are in fact obtaining the replication of the actual risk,

$$\hat{R}_{act}^*(b) = t(\omega, V_{out}^*(b)). \quad (2.52)$$

The bootstrap replication of the averaged actual risk can be obtained by just taking the average of $\hat{R}_{act}^*(b)$, and scatter measurements can be readily obtained from the same histogram. A typical range for B in practical applications can be in (50, 2000) bootstrap resamples.

Paired Benchmarking of Actual Risk with Bootstrap Resampling

For giving a clear cut-off test allowing us to benchmark the significance of the performance differences between two different ML based predictive models, we use here the previously described bootstrap nonparametric resampling procedure. We present the operative procedure in two complementary stages: first, the bootstrap based characterization of the residuals of a single model is introduced, allowing the detailed statistical characterization of the figures of merit under analysis; and second, the performance comparison between two models is described from

paired bootstrap resampling, which allows us to control for the standard error of the estimates of the differential figures of merit, giving moderate standard errors and allowing to establish cut-off tests for model comparison purposes.

The characterization of the *pdf* of the residuals for a single model can be summarized as follows:

- Given the original residual vector for a given model, $\mathbf{e} = [e_1, e_2, \dots, e_n]$, where the actual risk is evaluated from these residuals, then B independent bootstrap resamples are built, $\mathbf{e}^*(1), \mathbf{e}^*(2), \dots, \mathbf{e}^*(B)$, each given by n data resampled with replacement from the original residual set.
- For each resample, the value of a given figure of merit \hat{R}_{act}^* is calculated, and used as an estimation of the figure of merit under study, by using operator t , this is,

$$\hat{R}_{act}^*(b) = t(\mathbf{e}^*(b)). \quad (2.53)$$

Note that \hat{R}_{act}^* can be given by any of the figures of merit previously described.

- A sample distribution is built for the replications of statistic $\hat{R}_{act}^*(b)$, which stands for an approximation to the actual distribution for statistic R_{act} , and it can be an estimation of either average or scatter statistical description of the figure of merit.

From the sample distribution of \hat{R}_{act}^* , the 95% CI can be obtained, and its empirical value belonging to this interval will allow us to assume that the empirical estimator does not present a significant bias due to overfitting.

The previous procedure can be readily modified in order to benchmark the performance of two different ML techniques (or the same technique with different settings), by using a paired bootstrap resampling, with the same resamples considered in the benchmarked models. The procedure can be summarized as follows:

1. The residuals or the figures of merit yielded by two different ML based models, \mathbf{r} and \mathbf{s} , are considered, given by $\mathbf{r} = [r_1, r_2, \dots, r_n]$, and $\mathbf{s} = [s_1, s_2, \dots, s_n]$ and the differential resamples are built for the magnitude increments of these figures of merit, this is, $\Delta = |\mathbf{r}| - |\mathbf{s}|$, hence the differential increment resamples are $\Delta^*(1), \Delta^*(2), \dots, \Delta^*(B)$.
2. From these resamples of the increments, the increment in performance measurement $\Delta \hat{R}_{act}^*(b)$ is calculated, to be used as an estimator of the populational figure of merit under study, this is,

$$\Delta \hat{R}_{act}^*(b) = t(\Delta^*(b)). \quad (2.54)$$

The normalized histogram of the incremental performance is built for the statistic under analysis, which represents an approximation of the actual distribution of its *pdf*.

In this work, when resampling two different ML models $model_1$ and $model_2$, results will be compared in terms of average and scatter measurements according to three different statistics, namely,

$$\Delta MAE = MAE(model_1) - MAE(model_2), \quad (2.55)$$

$$\Delta CI(model_i) = CI_{sup}(model_i) - CI_{inf}(model_i) \quad (2.56)$$

$$\Delta CI = \Delta CI(model_1) - \Delta CI(model_2), \quad (2.57)$$

$$\Delta CI_{sup} = CI_{sup}(model_1) - CI_{sup}(model_2), \quad (2.58)$$

where CI has been obtained for 95% confidence level, CI_{sup} (CI_{inf}) are the superior (inferior) CI limits. These statistics give a description not only in terms of the average magnitude of the error, but also in terms of its scatter. Given that inference-based closed forms for CI scatter measurements are often a mathematically complex problem, it comes clear that bootstrap resampling represents a useful approximation for making it possible.

This theoretical contribution is applied in several applications. For example, in Application 3.1 (Sec. 3.2) both actual and paired risk estimation with bootstrap resampling are computed to evaluate the statistically significant differences among the considered ML techniques. In Application 3.2 (Sec. 3.3) and Application 4.1 (Sec. 5.2), several estimation models are considered and benchmarked using the paired bootstrap resampling approach presented in this section.

Machine Learning for Promotional Decision-Making

3.1 Introduction

The current economic landscape, characterized by financial instability and the consequent changes in consumer behavior, is driving a transformation in retailer decision, bringing to a new and more aggressive promotional perspective [87]. As an example of this situation, the dramatic sales reduction of some products in Spain, which has led retailers in the industry to implement new approaches, such as the intense use of private label products, can be mentioned. In addition, it has been also searched to increase consumer's frequent purchases through promotional activities, such as promotional discounts, feature advertising, and promotional packs (e.g., "buy 3 and get 1 free") [87]. Therefore, sales promotions have become in recent years a fundamental tool for retailers' strategies, and the investment in this setting has highly increased in the marketing strategy, with percentage even above 50% [88]. The better understanding of the sales promotion dynamics has received growing attention from ML and data mining techniques, which are powerful tools to extract information from recorded examples [17].

Existing models for analyzing sales promotions effects can be classified into two separate groups. In the first group, namely theoretical models, consumer behavior is basically evaluated by considering a sociological and psychological perspective, whilst in the second group of empirical models, promotional structures based on empirical information extracted from historical databases are usually built. Within that last group, the efforts have been focused during the last decades on the understanding of sales promotion dynamics based on classical statistical analysis methods, and more recent works are concentrated towards ML and data mining techniques, as powerful tools to extract information from existing recorded data [18, 19]. ML techniques aim to

find recurring patterns, trends, or rules, which can explain the data behavior in a given context, and then allows to extract new knowledge on the consumer behavior, to improve the performance of marketing operations. In particular, a vast amount of knowledge has been extracted from ML techniques, although not all the promotional behaviors have been studied and there is still room for further studies [15, 16, 17]. More specifically, operational problems arise in ML promotional modeling, when based on nonlinear estimation techniques, for evaluating and demonstrating working hypothesis [19, 20, 21, 22, 23, 24, 25]. First, conventional parametric tests are often not appropriate, because given the heavy tails and heteroscedasticity for the prediction residuals, Gaussianity is no longer a working property for them. Second, special attention has to be paid in order to be sure when working with hypothesis tests in terms of actual risk comparisons, and not of empirical risk comparisons, to avoid as much as possible the unaware presence of overfitting in the ML based models. And third, as an indirect consequence of not having a clear cut-off test, their results cannot always be easily compared across studies, even when they have been made on the same data set.

Therefore, the objective of the first application of this chapter, Application 3.1, is to propose an operative procedure for model diagnosis using ML techniques for promotional efficiency applications at store level. An empirical approach, based on ML techniques, is used for analyzing the sales dynamics for two representative databases with different promotional behavior, namely, a non-seasonal stable category (milk) and a heavily seasonal category (beer). Four well-known ML techniques with increasing complexity are benchmarked, specifically, k -NN, GRNN, MLP, and SVM. The nonparametric statistical tool based on the paired bootstrap resampling approach (see Sec. 2.5.2) is used for establishing a clear statistical comparison among them.

In addition, in Application 3.2, an operative and reliable analysis tool for promotional decision making based on retail aggregated data is also proposed. The main contribution from a digital signal and data processing viewpoint is the proposal of a new data-driven model based on a new set of indicators for evaluating the reliability and stability of a data model in terms of multidimensional feature space rather than a single merit figure for the model (e.g., the mean absolute error). These indicators allow the user to identify the uncertainty of different feature space regions, for example, unusual promotion conditions. Using the statistical processing available, the performance of different methods and different feature spaces is studied. The use of aggregate data in suitable conditions yields moderate and acceptable confidence intervals in these feature spaces.

3.2 Application 3.1: Promotional Efficiency at Store Level

3.2.1 Introduction

Though many definitions have been published for the term sales promotion [88, 89, 90], none of them are generally accepted, but general consensus suggests that sales promotions consist basically in short-time sales incentives [88, 89]. For instance, the American Marketing Association defines sales promotion as a media and non media marketing pressure applied for a predetermined, limited period of time in order to stimulate trial, increase consumer demand, or improve product availability [91]. Some researches [88] consider sales promotion not just a marketing element, but instead included within the strategic activity undertaken by the company. The sales promotion strategy adopted by the grocery retailer must be consistent with the general pricing policy. In fact, some strategic aspects of the retailer's pricing policy cover certain considerations related to the appropriateness of the use of promotions and discounts. For this reason, when under certain circumstances the use of deals and discounts are considered adequate, the specific discount rate must be determined attending to timing, frequency, and magnitude of the promotional discounts, [92, 93, 94, 95, 96].

Some studies suggest that the pricing policy adopted by retailers is influenced by many diverse aspects [97], among them factors related to the industry, the company itself, and other elements derived from the competitive situation and consumer demand. When referring to a specific activity of sales promotions, such as price promotion, it is important to make reference to the deal effect curve DEC, which shows the representation of actual sales volume against price discounts applied during a certain period. Hence, the DEC shows pricing and volumes, and depicts pricing promotions effects over different products, such as private label and/or normal brands. Effects illustrated by the DEC can be basically grouped into three categories:

1. The first category is related to direct discount effects. Two fundamental effects can be showed as far as this category is concern, namely, threshold and saturation. Threshold stands for the minimum discount that has to be applied to ignite sales growth [98], while saturation effect could be defined as the discount level that does not generate additional sales. This second effect can be justified either from the maximum number of product units that consumers can stock at home (especially with perishables products) [15], or from the consumer perception of discount itself, which has been shown to be lower than the real discount [99].
2. A second category relates to the cross-effect generated from other products promotions. The cross-effects appear when other brands and categories promotion indirectly imply variation on the volume sold of a certain product. This variation could be different depending on the value assigned by consumer to the promoted brand (providing a

much higher effect as the value perceived by the brand is higher) [23], and also depending on whether simultaneous promotions are inside the same category or substitutive products [100].

3. And finally, a third category is related to the number of concurrent promotions and their special characteristics, especially the different media used for the promotion, what implies different outcome (i.e. combination of an especial exhibition and temporary discount provides additional effectiveness to the promotion).

An increasing attention is being paid to the potential explanatory possibilities of ML techniques in promotional effectiveness. In [20], an algorithm based on historical transactions data, and yielding self and cross-effects for promotional sales prediction, was presented to estimate the promotions sales profits in retail sales and other business applications. In [25], Rough Sets and SVM techniques were used to establish a pricing model based on hedonic price improving prediction capabilities. In [19, 23, 24], sales promotion was modeled by means of semiparametric regression and semiparametric SVM, with no further comparison to other possibly relevant ML techniques, partly due to the lack of a suitable cut-off test, capable of dealing with non Gaussian, heteroscedastic prediction residuals, and actual risk comparisons.

ML Techniques for Promotional Sales Modeling

ML techniques have emerged as powerful tools to extract relevant quantitative information [17, 18]. Two different types of regression methods have been mostly used in the sales promotion literature to analyze the sales response to price promotions discounts: parametric regression and nonparametric regression. Parametric regression assumes a certain functional form underlying the data, namely linear, exponential, or logarithmic. The simplest parametric regression model is the linear model, where the parameters can be easily estimated using ordinary least squares, assuming the presence of additive, uncorrelated, and Gaussian white noise. However, in the presence of heteroscedasticity, generalized least squares methods are more appropriate [101]. In addition, maximum likelihood models assume a given statistical distribution linking the parameters and the data [102]. Nonparametric regression does not assume any a priori functional form, but it rather relies on approximating the observations locally. Examples of nonparametric methods are spline regression, k -NN, or kernel estimators [102]. The main advantages of nonparametric methods are flexibility and consistency, which are established under much more general conditions than for parametric modeling.

General data model for promotional sales. In order to support the model architecture that is capable of learning from the relationships between inputs (\mathbf{x} , column vector) and outputs (y), it is required a finite number of paired observations. In sales promotion modeling, the input

Table 3.1: *Products under analysis in the milk and beer product category.*

Model	Product (Milk)	Product (Beer)
Model 1	Asturiana	Amstel 25 cl x 6
Model 2	Ato	Amstel 33 cl
Model 3	Private brand	Bavaria 33 cl
Model 4	Pascual Calcio	Cruzcampo 33 cl
Model 5	Pascual Clásica	Estrella 25 cl x 6
Model 6	Puleva Calcio	Estrella 25 cl x 12
Model 7	-	Estrella 33 cl
Model 8	-	Heineken 25 x 6
Model 9	-	Private brand 33 cl
Model 10	-	San Miguel 25 cl x 6
Model 11	-	San Miguel 25 cl x 12
Model 12	-	Voll Damm 25 cl x 6
Model 13	-	Xibeca 25 cl x 6
Model 14	-	Xibeca 33 cl

pattern may consist of information about price changes and promotion characteristics, whereas the output would correspond to the number of sold units for a given product. The model $f(\cdot)$ for the relation $y = f(\mathbf{x})$, has been mainly estimated in the marketing research literature by using two different families of regression methods. Regarding to the first of them, in parametric methods, it is assumed a previously known shape or structure for functional relation $f(\cdot)$. In this case, the functional is often defined by a simple relationship (linear), while the nonparametric method does not assume any prior structure in terms of data model, instead, it is built the estimated relationship based on kernels (for instance, the Gaussian kernel) [4].

3.2.2 Database

Two real databases from the milk and beer product categories were analyzed. These two categories represent products with different promotional dynamics, in particular, milk is a daily used product, while beer is a highly seasonal product. We used the information extracted from both product categories, obtained from digital archives of sold units in the same retailer (supermarket) during one year, excluding weekends. Up to 304 examples (samples) were available for each category, corresponding to the days when transactions were recorded in the supermarket. Information was aggregated into 43 weeks, to avoid weekly seasonality effects that were clearly present in the data.

On the one hand, the milk category database was studied separately from beer products, in order to compare daily products. Hence, 6 brands were analyzed within this product category, corresponding to 6 different promotional models, as indicated in Table 3.1. On the other hand,

the beer database was assembled, as a reference of a strong seasonal product along the year. This separate structure of databases allowed the benchmark between categories with different behavior, as well as a benchmark among all models inside each category in order to compare with a strongly along the year seasonal product. In this second category, the promotional behavior of 14 different models were analyzed, as shown in Table 3.1. For some brands, different formats were considered (i.e. for Amstel beer a distinction was made between the 33 cl. can and the 25 cl. 6 units pack). In all models, exogenous variables were given by the set of price indices (PI) for the different items in each category. The price index for a given product is given by

$$PI(i, t) = \frac{P_{prom}(i, t)}{P_{reg}(i, t)} \quad (3.1)$$

where $PI(i, t)$ is the price index of product i at week t , $P_{reg}(i, t)$; and $P_{prom}(i, t)$ are the regular and promotional prices of product i at week t , respectively. Hence, the price index gives the relative variation between the promotional price and the regular price, and its value is 1 whenever both are equal. This index allows a clearer comparison of the magnitude of the discounts, and so it is often considered as a more informative exogenous variable than the promotional price. In addition to the price index of all the competing brands in a category for each model, other exogenous variables were also considered. First, a variable for direct discount (DD) (or equivalently, price reduction) was considered as a dichotomic variable (1 for existing direct discount and 0 otherwise). Second, a pre-processing algorithm was used for distinguishing between two possible seasonality-dependent behavior, by splitting each database into two possible periods. The first period was identified with 0 and the second with 1, allowing a natural way either for identifying the low from the high season, or for canceling its effect in the model. This dichotomic variable was called baseline (BL). Baseline sales is a key concept in marketing research and it is typically defined as the sales of a given product when there are neither marketing promotions for this product, nor promotions for other interacting products [103, 104]. Graphical representation of these assessments can be found in Fig. 3.1 where the PI for each model in both product categories, as well as the weekly sold units, are shown.

The promotional models for both databases share some characteristics, namely, the input sample is given by a combination of both price indices and dichotomic variables, and also the output of each model is given by the sold units for that particular product in the database. Hence, the promotional model can be expressed as:

$$y(i, t) = f(x^M(i, t), x^D(i, t), BL(t)) \quad (3.2)$$

where $y(i, t)$ is the number of sold units for product i during week t ; $x^M = [PI^1(i, t), \dots, PI^{n_m}(i, t)]^\top$ is a vector with the price indices of product i during week t , with $n_m = 6$ for milk database and $n_m = 14$ for beer database; $x^D(i, t)$ is the direct

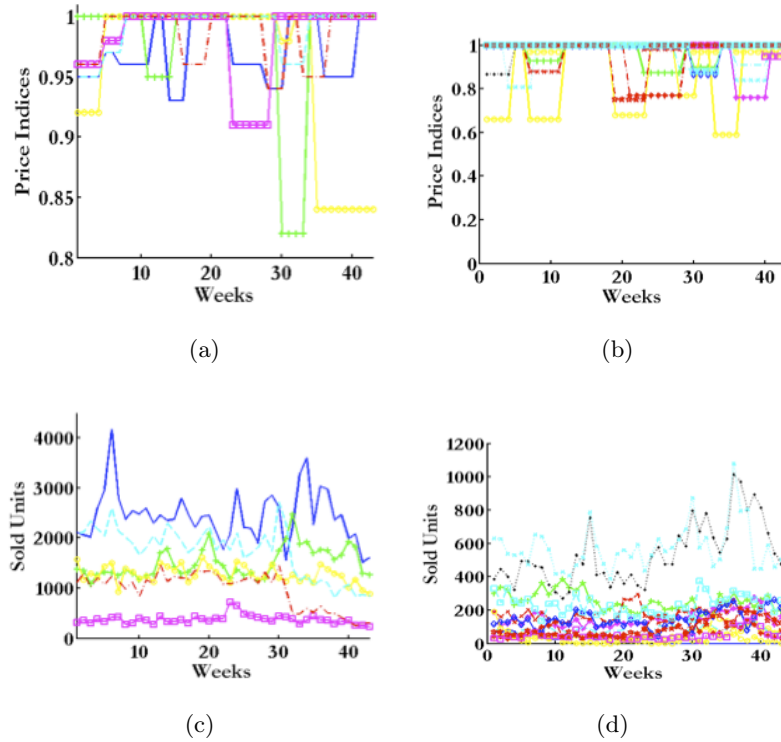


Figure 3.1: *Time evolution of price indices (a,b) and sold units (c,d) for mil database (a,c) and beer database (b,d).*

discount dichotomic variable for product i during week t ; and $BL(t)$ is the baseline for the dichotomic variable at week t .

3.2.3 Experiments and Results

Limitations of DEC Characterization with the Own-Effect

In order to verify whether a complex model is really required to analyze the existing data, the DEC was estimated by considering only each price index own-effect, and estimation was executed by calculating the average units of products sold as a function of the pricing index, without considering presence of simultaneous promotional effects by other competing or substitutive products.

Individual own-effect for models corresponding to milk category, and for beer category were obtained. In both cases DEC corresponds only to the effect, over each product, due to the discount applied, without taking into account any further interactions with other competitor products, which ended not being a fair approximation attending to results obtained. In many cases DEC shapes found could be explained according to direct effects such as threshold, saturation and price/demand standard elasticity, although other situations are also identified. In an attempt

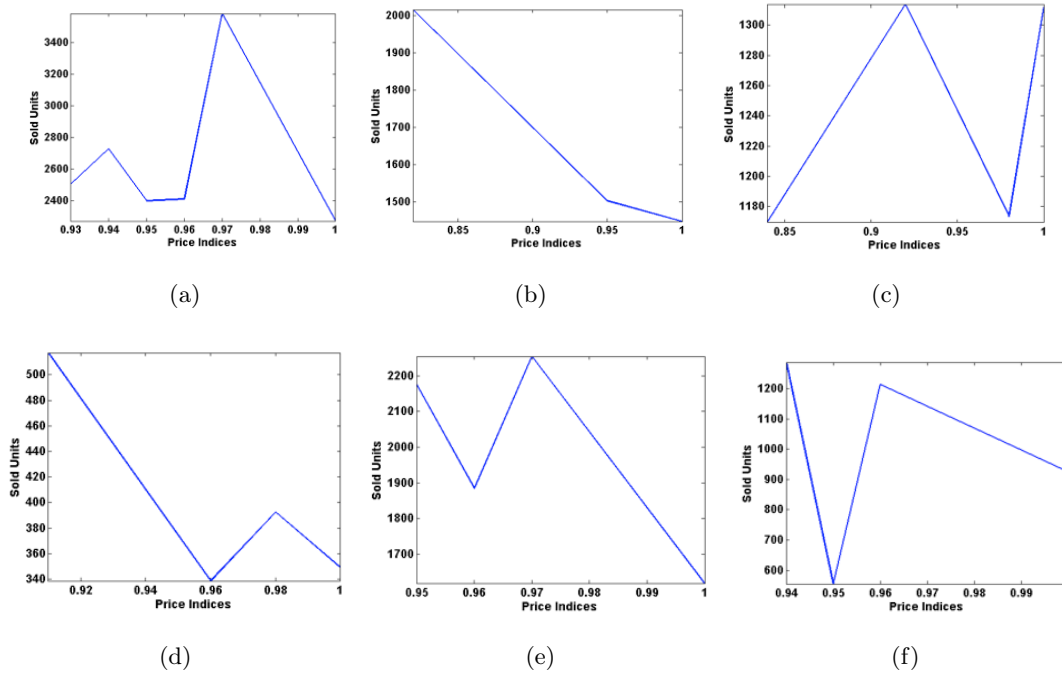


Figure 3.2: Averaged price index as a function of sold units for milk database. Models 1 to 6 correspond to panels (a) to (f).

to identify those effects in the individual DEC curves for milk category, it could be said that the largest elasticity effect can be found for the product presented in Fig. 3.2 (a), while the smallest one is detected in Fig. 3.2 (d). Threshold effects could be observed in Figs. 3.2 (c) and (d) at levels of discount of 2%, whilst the saturation effects could be detected in Figs. 3.2 (a), (c), (e) and (f) at levels of discounts of 3%, 8%, 3% and 4% respectively. Similar conclusions could be extracted from DEC simple estimations for the beer category.

Therefore, it seems reasonable to assume that, apart from threshold and saturation effects, other phenomena are taking place simultaneously, for instance, cross-price effects and differential interactions among promotional initiatives. Therefore, the DEC own-effect should be considered with caution, as it does not allow to identify or detect cross-effects in real data, more complex and sophisticated modeling techniques must be used. This is the motivation for applying ML techniques in order to detach and characterize existing simultaneous promotional effects.

Selection of Design Parameters for ML methods

The design parameters selection was just dependent on the ML method. In summary, four nonparametric regression/estimation techniques were used for comparison purposes: k -NN, GRNN, MLP and the linear and nonlinear ν -SVM. Fundamentals of the four techniques are

explained in Sec. 2.3.1. LOO was selected as the cross-validation approach due to the reduced number of available samples (see Sec. 2.5.1).

k -NN design parameter selection. As summarized in Sec. 2.3.1, the number k of neighbors considered for local-averaged estimation is the design parameter of this technique. Best value for k depends on the size and dimension of the database, and must be chosen so that the model is neither sensitive to atypical samples nor provides over-averaging, since both cases give poor quality estimations. For each model, parameter k is explored in the range (1, 40). Note that there is often a noticeable plateau, giving a stable working zone for the design parameter selection. In addition, it is worth to note that k tends not to be a low value, hence indicating that smoothing is necessary to yield better models with this technique and database.

GRNN design parameter selection. The implemented GRNN architecture is a multiple output scheme (Sec. 2.3.1) such that every output corresponds to one predictor variable (sold-units of a product in the database). The width σ of the symmetrical Gaussian kernel is the design parameter for the GRNN. Though from a mathematical point of view, outputs are uncoupled (they do not model cross-interactions among outputs), and hence the implemented architecture is equivalent to a set of independent individual models, the multiple output implementation has operational and computational advantages. Parameter σ was explored in the interval (0.01, 0.5).

MLP design parameter selection. The chosen MLP architecture (see Sec. 2.3.1 for details) has one hidden layer and multiple output so that optimization is simultaneously performed for all outputs, which are now coupled along with the design process. The design parameter to be selected was the number of neurons in the hidden layer, denoted as n_0 . Weights of the MLP were determined by applying the iterative Levenberg-Marquardt algorithm. To avoid overfitting an early-stopping procedure with holdout cross validation was performed.

SVM design parameter selection. As explained in Sec. 2.3.1, different free parameters have to be tuned in the ν -SVM. When working with linear kernels, two design parameters are necessary: $\nu \in (0,1)$ to control training errors and regularization parameter C ($C > 0$). The kernel width σ has to be also tuned when Gaussian kernels are considered. In the linear case, the (ν, C) space is explored. In the nonlinear approach, and to reduce the computational burden required by an exhaustive three-dimensional exploration, the following iterative procedure starting from an initial value of C was applied: (1) for a given C , the (ν, σ) space is explored; (2) pair (ν, σ) providing a minimum MAE is found; (3) with values of (ν, σ) obtained in previous step, parameter C is explored and the best conditional value is chosen; and (4) the obtained MAE is stored, and first and third steps are repeated until MAE becomes stable. The previous procedure was performed four times, each with a different initial value for C , with $C = [10, 50, 100, 1000]$.

Table 3.2: MAE for individual models in milk database. For each cell, empirical and bootstrap-averaged MAE (first row), and 95% CI (second row). Bold emphasizes the method with the best performance for each model.

	k-NN	GRNN	MLP	RBF SVM	Linear SVM
Model 1	366.44 365.50 [269.95,465.16]	366.44 365.50 [269.95,465.16]	320.75 320.56 [247.23,407.72]	322.15 320.68 [231.08,422.75]	389.43 388.00 [282.24,512.49]
Model 2	205.21 205.56 [156.94,257.97]	205.21 205.56 [156.94,257.97]	198.89 198.22 [154.57,244.11]	166.09 166.13 [123.03,214.02]	240.07 240.68 [185.46, 301.55]
Model 3	164.62 165.11 [131.26,199.64]	164.62 165.11 [131.26,199.64]	151.23 151.83 [122.15,182.68]	136.12 135.86 [109.63,165.34]	160.58 160.38 [124.98,199.50]
Model 4	63.27 63.18 [48.61,79.27]	63.27 63.18 [48.61,79.27]	72.41 72.85 [58.06,89.09]	50.64 50.41 [37.87,65.05]	71.73 71.72 [51.25,93.84]
Model 5	178.99 178.49 [131.93,226.97]	178.99 178.49 [131.93,226.97]	196.36 196.83 [151.58,248.20]	192.59 193. 03 [134.98,257.77]	380.79 380.89 [282.89,494.88]
Model 6	119.68 119.54 [92.13,148.40]	119.68 119.54 [92.13,148.40]	127.68 127.64 [99.38,157.79]	105.13 105.24 [80.24,132.32]	238.90 238.85 [178.34, 307.87]

Benchmarking Prediction Models

The first analysis to be made is the benchmarking among different ML techniques (k -NN, GRNN, MLP, linear and nonlinear SVM) in terms of LOO-based MAE, for each product and both databases. For the milk database, the promotional model considered 8 input variables, namely, 6 metric variables corresponding to the price indices of itself and of competitor items, and 2 dichotomic variables (direct discount indicator and seasonality). The number of model outputs was dependent on the technique, this is, 6 product models with 1 single output each for k -NN and ν -SVM, and one joint model with 6 outputs for GRNN and MLP. Table 3.2 shows the results of the empirical MAE (recall again that it has been obtained with LOO and averaging), together with the bootstrap-averaged MAE. Note that consistence between these two values is an indicator that generalization capabilities are properly quantified with the merit figure, whereas a reduced empirical MAE compared to the bootstrap-averaged will be an indication of overfitting. The 95% CI is also summarized, in order to give a nonparametric measurement of scatter for each method. Bold typeface emphasizes the best ML technique, in terms of averaged MAE. From the values in the table, we can observe that RBF SVM seems to be the best ML technique for Models 2, 3, and 4, whereas k -NN seems to be the best one for Model 5.

For the beer database, the promotional model considers 16 input variables (14 metric variables, and two dichotomic variables for direct discount and seasonality), with the same considerations as before for the number of outputs and models in each technique. Results in terms of MAE are shown in Table 3.3, where it can be observed that RBF SVM seems to be the best method for all the models, except for Models 2 and 7, where k -NN seems to be the best

Table 3.3: MAE for individual models in beer database. For each cell, empirical and bootstrap-averaged MAE (first row), and 95% CI (second row). Bold emphasizes the method with the best performance for each model.

	k-NN	GRNN	MLP	RBF SVM	Linear SVM
Model 1	7.85 7.85 [5.16,10.94]	11.85 11.90 [9.07,14.83]	13.86 13.75 [10.33,17.50]	6.98 6.96 [4.77,9.16]	12.93 12.92 [8.85, 17.34]
Model 2	33.77 33.61 [25.67,42.11]	47.56 47.43 [38.10,57.32]	49.76 49.42 [39.08,62.28]	34.52 34.53 [27.33,42.32]	48.50 48.57 [36.61,59.32]
Model 3	15.10 14.96 [11.48,19.38]	14.92 14.70 [10.94,19.35]	14.33 14.28 [10.49,18.60]	14.08 14.11 [10.48,18.67]	15.40 15.46 [11.53, 20.29]
Model 4	25.75 25.74 [19.59,32.09]	27.51 27.58 [22.34,33.32]	26.74 26.42 [20.15,34.00]	21.40 21.40 [16.36,27.01]	29.53 29.36 [22.81,36.04]
Model 5	12.21 12.05 [8.82,15.92]	13.67 13.67 [10.67,16.49]	13.62 13.54 [10.28,16.77]	10.27 10.25 [7.54,13.19]	12.85 12.87 [10.38, 15.31]
Model 6	31.10 30.89 [23.82,39.15]	32.33 32.15 [24.29,40.91]	31.59 31.65 [25.15,38.95]	26.82 26.86 [18.92,35.37]	32.21 32.41 [24.67,41.85]
Model 7	78.86 78.57 [61.13,99.64]	97.54 96.86 [72.26,123.36]	103.53 102.14 [72.98,136.63]	89.50 89.53 [64.13,118.31]	102.08 102.20 [75.92,130.58]
Model 8	36.90 36.92 [29.91,48.91]	36.19 36.19 [26.89,46.31]	47.42 47.15 [38.34,57.09]	34.59 34.71 [26.83,43.39]	36.03 36.25 [27.42, 46.24]
Model 9	20.77 20.64 [16.04,25.67]	32.93 32.85 [26.50,39.82]	28.92 28.70 [21.18,38.62]	19.58 19.62 [14.48,25.02]	31.36 31.43 [23.87,39.84]
Model 10	17.98 17.88 [11.54,26.99]	24.12 24.02 [16.44,34.14]	26.08 26.22 [18.29,35.32]	12.85 12.84 [8.39,19.15]	21.69 21.82 [13.09, 32.44]
Model 11	13.65 13.62 [9.74,17.73]	17.48 17.42 [12.51,23.41]	13.60 13.62 [9.86,18.06]	11.48 11.43 [7.92,15.18]	17.62 17.63 [13.29,22.93]
Model 12	109.58 109.69 [83.88,140.81]	102.62 102.00 [77.19,130.96]	146.30 146.07 [107.64,183.87]	96.85 97.06 [71.72,126.61]	99.46 99.85 [72.24,130.99]
Model 13	16.49 16.45 [12.88,20.21]	22.83 22.69 [18.21,27.91]	26.16 25.92 [18.84,34.56]	14.91 14.94 [11.47,18.78]	22.23 22.32 [16.96, 28.41]

scheme.

In general terms, it can be concluded that, for MAE as merit figure, RBF SVM is the technique with better performance. This advantage is more patent in the case of beer database products, and occasionally, k -NN yields better performance than RBF SVM. With respect to the remaining ML techniques, it is often complicated to benchmark in terms of averaged MAE. For instance, in the milk database, GRNN gives lower MAE than k -NN and MLP for Models 2 and 3, but not for the remaining products. Therefore, in order to give a clear cut-off test allowing the comparison, the next step is to use the proposed bootstrap paired test.

Table 3.4 shows the paired comparison of k -NN vs GRNN, k -NN vs RBF SVM, GRNN vs RBF SVM and linear SVM vs RBF SVM. Bootstrap resampling allows us to calculate the 3 different

Table 3.4: Paired bootstrap for milk Database using three statistics for MAE merit figure in each cell, namely, ΔMAE (first row), ΔCI (second row), and ΔCI_{sup} (third row), with mean and 95% for each. Bold emphasizes the technique comparison with statistically significant differences at 95%.

	k-NN vs GRNN	k-NN vs ν -SVM	GRNN vs RBF SVM	Linear vs RBF SVM
Model 1	-33.26 [-107.37,43.21] 150.41 [-453.69,749.42] -54.43 [-499.08,438.29]	45.14 [-6.46,106.76] 231.63 [-344.72,706.90] 8.06 [-247.29,163.40]	76.78 [20.21,133.73] 86.76 [-469.78,431.90] 158.71 [-304.17,300.46]	21.51 [-45.62,91.75] 90.12 [-176.88,535.20] 101.64 [-173.43,537.06]
Model 2	25.60 [0.09,54.63] -1.98 [-135.32,121.75] 21.94 [-106.26,194.54]	38.22 [10.13,68.02] 21.14 [-169.14,226.54] 59.74 [-13.31,234.87]	12.13 [1.04,23.53] 18.35 [-21.31,93.62] 29.16 [2.40,109.07]	80.98 [29.29,130.67] 162.70 [20.73,454.97] 195.66 [50.14,489.21]
Model 3	16.95 [-1.18,33.74] 141.97 [46.41,246.08] 92.05 [14.20,142.75]	27.40 [7.47,48.29] 159.73 [39.62,260.05] 95.50 [-22.02,180.99]	10.40 [2.19,19.08] 15.92 [-37.84,104.71] 4.50 [-33.02,39.79]	8.15 [-4.57,21.04] 14.68 [-40.09,74.93] 16.92 [-38.42,76.60]
Model 4	1.74 [-15.80,14.41] -12.40 [-129.01,144.81] -52.49 [-122.24,98.23]	13.64 [3.43,23.68] 67.71 [-11.58,130.99] 10.79 [-35.07,74.79]	15.49 [4.53,26.47] 82.00 [-13.82,122.43] 64.60 [-32.74,92.66]	65.14 [-24.80,115.85] 4.33 [-6.73,15.37] 66.58 [-22.95,114.04]
Model 5	-15.22 [-55.52,28.75] 70.56 [-142.65,255.19] 1.10 [-63.07,57.34]	-11.40 [-52.92,31.54] -41.69 [-417.52,251.53] -202.82 [-401.77,-3.48]	2.86 [-39.87,44.85] -110.44 [-353.20,148.94] -203.46 [-407.88,23.06]	226.28 [147.89,303.90] 268,40 [152.12,413.05] 275.29 [177.14,407.94]
Model 6	4.04 [-19.98,27.76] 46.78 [-106.24,163.55] 29.39 [-36.50,111.73]	7.28 [-21.99,36.11] 55.53 [-92.50,169.06] 13.35 [-24.32,60.57]	2.62 [-12.44,17.51] 9.87 [-112.51,104.06] -15.63 [-51.16,38.43]	192.47 [137.81,240.53] 21.37 [-134.94,221.37] 77.20 [-24.26,240.90]

statistics previously described in Sec. 2.5.2, namely, the difference of averaged MAE (ΔMAE , first row), the difference in the width of CI of the MAE distribution (ΔCI , second row), and the difference between the upper limits of the CI (ΔCI_{sup} , third row), for paired-benchmarked ML methods. Recall that the two last measurements give a quantification of the scatter, whereas the first one gives a quantification of centering. Consistently with both conventional statistics, we can say that the performances of 2 methods are statistically different whenever the CI of the increment of the statistic does not overlap the zero level. Hence, in terms of CI limits, in the case both limits were simultaneously negative, it will indicate that the first technique significantly outperforms the second; both limits with positive values will indicate that the second technique significantly outperforms the first one. Trivially, a negative limit together with a positive limit indicates that no significant difference can be given to any of the compared methods.

Table 3.4 presents the results of analysis for dairy products using the 3 statistics. This structure of data studied, i.e. generating and analyzing the 3 statistics for paired comparisons and statistical testing, was extended throughout this research. However, the results obtained for the 2 statistics related to the scattering were consistently equivalent, and therefore, the remaining

Table 3.5: Paired bootstrap for beer Database using three statistics for MAE merit figure in each cell, namely, ΔMAE (first row), ΔCI (second row), and $\Delta CI_{s,up}$ (third row), with mean and 95% for each. Bold emphasizes the technique comparison with statistically significant differences at 95%.

	k-NN vs GRNN	k-NN vs ν -SVM	GRNN vs ν -SVM	Linear vs RBF SVM
Model 1	-4.07 [-6.88,-1.01] -2.42 [-8.79,12.76]	0.85 [-0.61,2.43] -6.73 [-13.01,7.88]	4.89 [2.08,7.85] -6.53 [-10.03,5.16]	5.93 [1.97,10.26] 19.62 [8.29,31.29]
Model 2	-13.51 [-24.07,-4.29] -11.33 [-51.85,19.99]	-0.75 [-4.40,2.89] -13.20 [-24.47,-1.24]	13.01 [3.43,23.07] 0.68 [-26.16,38.10]	14.04 [4.16,24.88] 32.36 [-5.94,67.87]
Model 3	0.27 [-3.22,4.04] -12.19 [-20.08,19.58]	1.00 [-2.06,4.11] -12.04 [-26.49,17.23]	0.87 [-1.43,3.27] -1.61 [-17.07,10.44]	1.32 [-1.57,4.37] 3.89 [-5.26,14.04]
Model 4	-1.78 [-7.06,3.04] 8.76 [-17.53,20.85]	4.30 [0.96,7.34] 0.91 [-23.05,12.03]	6.08 [0.92,11.16] -7.07 [-27.73,16.90]	10.72 [-6.30,50.22] 8.29 [2.47,14.34]
Model 5	-1.16 [-4.43,1.71] 7.02 [-1.55,17.34]	1.88 [0.12,3.75] 0.68 [-10.08,14.4]	3.36 [0.55,5.99] -6.78 [-13.52,1.50]	2.62 [0.43, 4.70] -5.89 [-17.70,4.58]
Model 6	-0.93 [-9.70,7.70] -9.67 [-44.34,20.41]	4.26 [1.82,6.93] -3.32 [-41.07,24.58]	5.50 [-4.14,14.83] 3.35 [-33.54,31.77]	5.41 [-3.84,14.47] 4.52 [-30.69,64.74]
Model 7	-18.25 [-43.59,3.21] -76.33 [-171.02,31.78]	-10.67 [-35.21,11.45] -47.81 [-261.81, 63.56]	7.94 [-5.48,20.40] 7.89 [-127.94,118.44]	12.81 [-0.60,25.69] 18.64 [-66.96,81.47]
Model 8	0.83 [-6.60,8.17] -13.11 [-46.22,13.68]	2.25 [-2.97,6.98] -19.78 [-44.14,-3.42]	1.61 [-3.86,7.24] -8.97 [-32.16,16.40]	1.51 [-5.05, 7.98] 17.25 [-0.79,32.86]
Model 9	-12.14 [-19.99,-4.32] -23.08 [-42.88,0.98]	1.21 [-1.28,3.70] -1.11 [-15.95,6.42]	13.30 [6.45,20.67] 17.54 [-4.85,41.78]	11.81 [4.23,19.70] 32.49 [7.70,55.29]
Model 10	-6.19 [-13.24,0.53] -11.69 [-46.93,17.99]	5.10 [0.69,10.34] -7.95 [-15.95,6.43]	11.17 [5.49,17.63] 5.98 [-20.11,37.46]	55.21 [20.70,70.61] 8.93 [2.68,16.33]
Model 11	-3.65 [-8.58,0.27] -14.88 [-43.21,1.57]	2.19 [-0.73,5.30] 0.66 [-13.85,13.62]	6.03 [2.68,9.41] 16.23 [1.44,33.81]	6.15 [3.03,9.72] 10.42 [2.65,18.40]
Model 12	6.47 [-20.01,30.79] 82.44 [-153.18,146.97]	13.04 [-18.12,44.29] 15.96 [-174.49,109.52]	5.73 [-1.99,14.50] -31.25 [-126.57,16.92]	2.45 [-12.58,18.47] 8.74 [-91.51, 71.80]
Model 13	-6.31 [-11.29,-1.8] -18.75 [-36.99,9.57]	1.62 [-2.11,5.07] -29.97 [-44.63,-3.48]	7.90 [2.40,13.23] -11.61 [-36.42,18.84]	7.29 [1.84,13.06] 25.72 [-4.74,56.5]
Model 14	-0.71 [-8.51,6.23] 0.24 [-30.71, 14.10]	3.67 [-2.24,9.36] -20.96 [-81.49,-4.77]	4.83 [-1.04,10.46] -17.78 [-51.68,5.46]	12.81 [5.37,21.23] 38.62 [7.00,52.91]

results of this work will show only the first one (ΔCI). For this particular case, from the third column in Table 3.4, it can be concluded that it is better to estimate the number of sold units with RBF SVM for Models 1, 2, 3, and 4. However, in terms of ΔCI , there are no significant differences between GRNN and RBF SVM, and all histograms are centered on zero. Models 2 and 3 correspond to ATO brand and distributor brand, respectively. These two models gave significantly better performance when using RBF SVM than when using k -NN or GRNN in terms of averaged MAE, but there were no differences in terms of scatter merit figures.

When comparing the paired bootstrap test for analyzing the differences between k -NN and RBF SVM, the last one was significantly better for Models 2, 3, and 4. It can be also observed that in Model 5, the upper limit of CI is lower with k -NN scheme. Regarding linear and RBF SVM comparison, it can be concluded that RBF SVM yields better performance for Models 2, 5 and 6. This conclusion suggests that is better to use a nonlinear approximation to characterize the promotional efficiency in dairy product.

As a summary, for milk category, RBF SVM results performed better than k -NN in some models, and no significant difference was obtained in the rest of the cases, in terms of ΔMAE . GRNN performance mainly overcame k -NN, while ν -SVM, in general terms, also overcame GRNN. So, as a key result it could be mentioned that for milk category RBF SVM was the best performing method.

Paired bootstrap tests for beer database are shown in Table 3.5. When comparing k -NN vs GRNN, there are significant performance differences in ΔMAE for Models 1, 2, 9, and 13, k -NN yielding significantly better quality for the estimation. For the distributor brand (Model 9), the upper lower of CI was significantly lower when using k -NN. When comparing k -NN vs RBF SVM, in terms of MAE it was better to use RBF SVM for Model 4, 5, and 6, however, the scatter was lower when designing the models with k -NN, specifically, in Models 2, 8, 13, and 14, both for ΔCI . When comparing GRNN vs RBF SVM, it can be said that RBF SVM yielded a significantly better MAE for Models 1, 2, 4, 5, 9, 11, and 13. In terms of scatter, ΔCI was only significantly different for Model 11. Regarding linear and nonlinear SVM, it can be concluded that nonlinear approach performs better for most of the products.

As a summary conclusion of this experiment, we could state that using MAE merit figure, in absolute terms, RBF SVM method provided a better performance, although occasionally k -NN overcame SVM in the case of beer database, while in some cases it is GRNN the preferred method for milk database.

3.2.4 Discussion and Conclusions

In this study, two separate types of conclusions can be distinguished. First, those ones related to the evaluation of the different ML methods, and secondly, those related to promotion of a specific product. From a marketing point of view, it has been evidenced that it is essential to better understand not only the consumer behavior in terms of their response to price deals, but also an in depth study of the adequate methodologies. Thus, the tendency is to evolve to nonparametric regression methods, which allow more flexibility and a higher ability to adapt to the specific promotional features. This is really important in the scenario of the present study, where two databases corresponding to different food product categories with specific characteristics. Milk is a daily used product, while beer has a high level of seasonality. On

the other hand, milk is consumed by a higher range of consumers segments, whereas beer is consumed by adults.

Main contribution of this work is the proposal of an operative method to evaluate the promotional efficiency, based on ML, as a valid method to analyze the multiple and simultaneous effects coexisting in promotional activities in retail markets when using real-world data. Justification for the use of ML tools is based on the fact that real data are subject to a large number of factors (namely, consumer idiosyncratic behavior, multiple simultaneous promotions in the same category, promotions in complementary/ substitute products and categories, and even consumer share of wallet), and hence simple models would not be able to trace all concurrent events or even extract complex multivariable characteristics.

As far as method evaluation is concerned, initial results showed that very often it comes complicated to identify significant differences in the model quality for the ML techniques presented (k -NN, GRNN, MLP and SVM). Final results showed that RBF SVM presented a significant better performance, followed by k -NN and GRNN, for milk category. For beer category, results were also better in general terms for RBF SVM, although in some cases a better result was obtained using k -NN.

3.3 Application 3.2: Promotional Efficiency at Chain Level

3.3.1 Introduction

From a retail manager's viewpoint, sales forecasting is essential not only to set the right pricing for an individual product [105] but also to define the promotional structure that maximizes benefits within a category as a whole [106]. The same rationale applies to individual customer behavior with regard to the total impact of a certain promotional strategy [106, 107, 108, 109, 110]. As a consequence, promotional models built on market-level data are considered as the best suited to describe the market behavior. Executive decisions are mainly based on this kind of information, especially for those retail chains accounting for a significant market share. Although it is evident that aggregated retail sales forecasting could potentially improve store sales prognosis [111], nevertheless, many authors have warned against the biasing risk during the aggregation process [112].

For a decision-making tool to be an efficient instrument for promotional retail management, it must be designed to be operative and reliable. To be operative, the retail management tool should be able to handle data models that: (1) can be better described time series (TS) dynamics, static paradigms, or even by both; and (2) can be better represented by linear or by nonlinear dynamics. To be reliable, the tool must be more robust when working with aggregated data than working with store level data, but also must ensure an adequate aggregation process. We

describe next relevant data aggregation precedents and summarize conventional TS dynamic for promotional sales models. Static DEC and learning-based nonlinear methods were described previously in Application 3.1, thus, we avoid to repeat them in this application.

Data aggregation at chain level. Previous research has considered three levels of aggregation: store, chain, and market levels. At the *store-level*, data can characterize consumers' behavior (by considering buying habits such as products, and units to evaluate loyalty and churn rates), as well as brand or product sales (by aggregating sales). Household information for each product category can also be used to analyze the individual brand sales behavior and pricing effects can also be analyzed [113]. Further aggregation at *chain-level*, or even at *market-level*, integrates the information for brands or categories to provide accumulative effects [114].

According to published research analysis, each level of aggregation may introduce bias, which depends to a great extent on the aggregation method, thus limiting the generalization capabilities of the forecasting model. In [115], the authors analyze bias effect by comparing sales estimates at both store and chain level, and conclude that bias may be related to heterogeneous marketing strategies within stores. The authors also note that relevant information, such as marketing strategies followed by competitive retailers, is not reported or registered through scanner datasets. Other studies use different approaches to address model heterogeneity and bias among stores. For example, authors in [116] proposes a random coefficient demand model to avoid bias when data aggregated across stores with heterogeneous promotional activity are considered. However, bias may not be fully removed due to substitutive effects, competing products and heterogeneity; therefore, in the current study, we followed the methodology in [117], in which bias can be mitigated by aggregating data across stores with homogeneous marketing activities.

TS for promotional modeling. Promotional activities typically exhibit a strong temporal dependence, which suggests that certain models taking into account temporal variations could yield better results than static DEC. In this setting, the statistically well-founded TS analysis, has received a great deal of attention in the last decade of the twenty-first century, due to the vast amount of data available from electronic records and media (e.g., scanner data), which allows both the cross-sectional and longitudinal analyses [118]. Researchers have used TS techniques for forecasting marketing variables and for evaluating specific situations [119]. New tools based on TS have proliferated in recent years to support general decision-making and especially in marketing activities [118]. For example, autoregressive moving-average (ARMA) modeling provides a well-developed general framework to analyze time series. It can be further extended to take into account exogenous variables (so-called ARMAX models) to improve their predictive capabilities. A multivariate version of ARMA models, the vector ARMA, allows adjusting models in which the dependent variable can be explained by multiple TS [119].

In this Thesis, we propose an operative analysis tool for promotional decision making based on

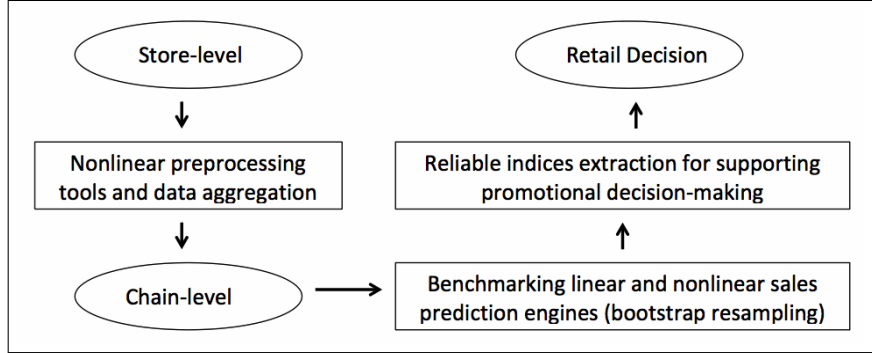


Figure 3.3: *Schematic of the proposed chain-level analysis.*

retail aggregated data. Using the statistical processing available, we can study the performance of different methods and different feature spaces.

3.3.2 General Forecasting Promotional Model

A three stages chain-level analysis as shown in Fig. 3.3 is proposed. First, we use a signal preprocessing method to aggregate data from store-level to chain-level, based on *a-priori* considerations and simple morphological analysis. Second, we propose a generic promotional sales forecasting tool that can account for static and time-varying dynamics models in a given product, while maintaining a simple and compact mathematical form. Decisions about different plausible models are determined by their comparative benchmarking by means of nonparametric resampling statistical tests (formally introduced in Sec. 2.5.2). Finally, the new proposed statistical indices are defined in the feature space and calculated for each product using resampling techniques.

We will generally consider data available at discrete time t , mostly consisting of prices and sold units in a weekly time period. Accordingly, $P_{i,k}(t)$ ($s_{i,k}(t)$) denotes the price (the number of sold units) for the i -th product at store k during week t , where $i \in \{1, \dots, I\}$, $k \in \{1, \dots, K\}$ and $t \in \{1, \dots, T\}$, with I, K and T being the total number of products, stores and weeks, respectively. Recall that $P_i(t)$ represents the price proposed by headquarters (HQ), which should be identical for the same product and week in all stores, however, day-by-day knowledge shows that prices are often different at each store due to promotional local decisions. This variability may be related to human errors during scanning process at cashier, special discounts applied due to damaged items, errors in the information systems, or even changes in prices due to local strategies. Store and central prices can be related by the following expression,

$$P_{i,k}(t) = P_i(t) + X_{i,k}(t) \quad (3.3)$$

where X is an uncertainty term.

Because we are interested in decision making according to central prices, we will need to approximate them by an adequate estimation operator Φ , this is,

$$\hat{P}_i(t) = \Phi\{P_{i,k}(t)\} \quad (3.4)$$

In contrast, sold units can be readily aggregated at the chain-level ($S_i(t)$) by the accumulative sum across stores, that is,

$$S_i(t) = \sum_{k=1}^K s_{i,k}(t) \quad (3.5)$$

After data preprocessing, a general forecasting model for the i -th product can be written as

$$S_i(t) = \hat{S}_i(t) + e_i(t) \quad (3.6)$$

$$\hat{S}_i(t) = F(\Theta_i\{\hat{P}_i(t)\}, \Xi_i\{S_i(t)\}) \quad (3.7)$$

where $\hat{S}_i(t)$ are the forecasted values of aggregated sold units at time t for the i -th product; e_i are the model residuals; operator F stands for the method used for estimation, such as DEC analysis, linear TS, or nonlinear statistical learning algorithms; and Θ_i, Ξ_i , denote the features extracted from prices and sold units series for the i -th product. Note that operator F gives a data description in the so-called *feature space*, defined by Θ_i and Ξ_i feature (column) vectors, which can be concatenated in a simple feature vector given by $\Psi_i = [\Theta_i^\top, \Xi_i^\top]^\top$. In the following, we will use both $\hat{S}_i(t)$ and $\hat{S}_i(\Psi_i(t))$ to denote the estimated number of sold units at time t for the i -th product.

The DEC model with just self-product effects can be readily expressed by $\Psi_i(t) = [\hat{P}_i(t)]$, as follows,

$$\hat{S}_i(\hat{P}_i) = E\{S_i(t) | \hat{P}_i(t) \in \varepsilon(\hat{P}_i)\} \quad (3.8)$$

The expectation operator is used to smooth the number of sold units with respect to observed pairs of price and sold units within a neighborhood of a given price, $\varepsilon(\hat{P}_i)$. Note that operator F in Eq. (3.7) is given by the price expectation within the \hat{P}_i neighborhood.

$AR(p)$ and $ARx(p, q)$ models, with p autorregresive terms and $q + 1$ exogenous input terms, can be written as follows,

$$\hat{S}_i(t) = \sum_{r=1}^p \phi_r S_i(t-r) \quad (3.9)$$

$$\hat{S}_i(t) = \sum_{r=1}^p \phi_r S_i(t-r) + \sum_{j=0}^q \theta_j \hat{P}_i(t-j) \quad (3.10)$$

where ϕ_t, θ_j are the model parameters [120, 121], and Eq. (3.7) is readily adapted by generating the following feature spaces,

$$\Theta_i\{\hat{P}_i(t)\} = [\hat{P}_i(t), \dots, \hat{P}_i(t-q)]^\top \quad (3.11)$$

$$\Xi_i\{S_i(t)\} = [S_i(t-1), \dots, S_i(t-p)]^\top \quad (3.12)$$

Table 3.6: *Price level, brand and regular and promotional sold units (%) for each product.*

Product	Price level	Brand	Regular	Promotional
			Sold Units (%)	Sold Units (%)
Product 1	0.3426	Brand 1	43.82 %	51.99 %
Product 2	0.2860	Brand 1	43.82 %	51.99 %
Product 3	0.2489	Brand 2	7.62 %	12.20 %
Product 4	0.2181	Brand 2	7.62 %	12.20 %
Product 6	0.1810	Brand 3	5.22 %	5.38 %
Product 5	0.1522	Brand 3	5.22 %	5.38 %
Product 7	0.1455	Brand 4	1.86 %	1.62 %
Product 10	0.1372	Private label	17.19 %	8.74 %
Product 8	0.1004	Brand 5	2.24 %	0.00 %
Product 9	0.0836	Brand 5	2.24 %	0.00 %

Thus, we can simply use $F(\Theta_i, \Xi_i) = \phi^\top \Xi_i + \theta^\top \Theta_i$ for the ARx model accounting for past prices as exogenous variables, and $F(\Theta_i, \Xi_i) = \phi^\top \Xi_i$ for the AR model, which is only built on the self-dynamics of the observed time series without information about prices.

In addition, nonlinear data models can readily be taken into account with the model nomenclature in Eq. (3.7). For the current study, k -NN technique is used as a nonlinear method for promotional sales forecasting, due to its extreme simplicity and acceptable performance in many applications. The k -NN estimator in t_0 is a nonparametric procedure that just consider the k nearest data to $\Psi_i(t_0)$, according to a given similarity or distance measurement [49], where k has to be previously fixed during the design procedure. Conventional distances are L_1 and L_2 norms, though different measurements have been proposed according to the nature of the data [18].

The k -NN estimator assumes that data close in the feature space Ψ provide similar values for the independent variable. Therefore, to estimate the number of sold units at any time t_0 , $\hat{S}_i(t_0)$, the k -NN estimator uses a local neighborhood $\kappa(t_0)$ to provide the estimation as

$$\hat{S}_i(t_0) = F_{\kappa(t_0)}\{S_i(t)/t \in \kappa(t_0)\} \quad (3.13)$$

where $F_{\kappa(t_0)}$ is the weighted average operator that depends on distance and parameter k , and it is given by: $F_{\kappa(t_0)} = \frac{\sum_{l=1}^k w_l S_i(t_l)}{\sum_{l=1}^k w_l}$, where $w_l = 1/d_l$ depends on the distance to the l -th nearest neighbor (d_l).

3.3.3 Database

Our database contains the weekly consolidated information from all electronically recorded data from scanners at cash registers for a Spanish store chain. The information is from 118 stores ($K = 118$), during 105 weeks ($T = 105$) between 2008 and 2009 for 10 products ($I = 10$).

Table 3.7: *Top ten retail stores in terms of sold units.*

Stores	Regular Sold Units (%)	Promotional Sold Units (%)
Store 1	2.72 %	2.61 %
Store 2	2.43 %	2.07 %
Store 3	2.16 %	1.95 %
Store 4	2.02 %	1.88 %
Store 5	1.80 %	1.20 %
Store 6	1.71 %	1.16 %
Store 7	1.65 %	1.75 %
Store 8	1.60 %	2.35 %
Store 9	1.59 %	1.37 %
Store 10	1.56 %	1.83 %

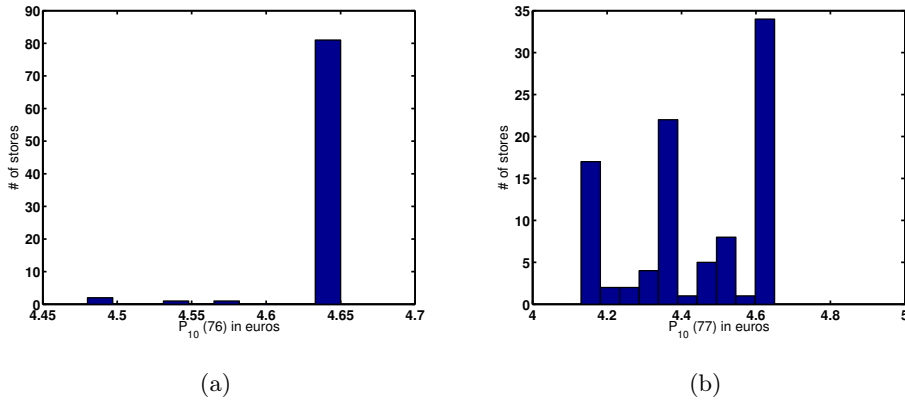


Figure 3.4: *Number of stores with a given price for Product 10 during two consecutive weeks, $t = 76$ (a) and $t = 77$ (b).*

We selected *Laundry detergent* as the best category to be analyzed in this study for several reasons: (1) it is an easily storable product with almost no expiration date; (2) almost all customers buy products of this category, and it is considered a basic household product; and (3) it was one of the largest products in the database in terms of sales. We assembled a database consisting of six brands in this category, including a private label (Table 3.6), and sold units were almost equal across stores (Table 3.7). The largest store in terms of sold units accounted only for the 2.72% of the total sales; however, this scenario is adequate for aggregation purposes.

Promotional activities are carried out by HQ, which means that prices are assumed to be identical for each product in every single store. Consequently, chain-level decisions and global strategies were considered as the main source of promotional activity, rather than store-level marketing strategies. However, database information showed remarkable variability in terms of pricing being applied across stores (Fig. 3.4), which reveals that real databases always incorporate

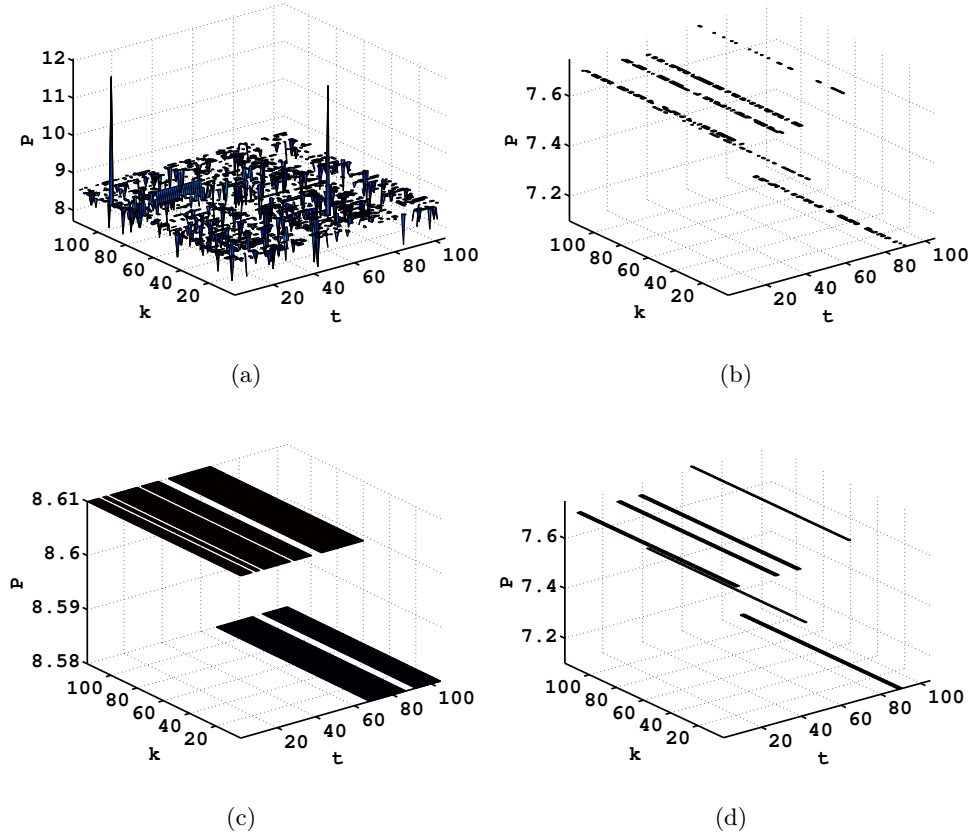


Figure 3.5: Prices for Product $i = 1$, where central prices are stated as regular (a,c) or promotional (b,d). The prices before (a,b) and after (c,d) preprocessing are also shown.

uncontrolled effects on actual pricing, regardless how strict a HQ's pricing is. As an example, Fig. 3.4 represents bar graphs for prices versus stores for one product in our database. These graphs demonstrate that the existence of a well-known statistical distribution shape is not a sustainable assumption. The high variability shown in these graphs suggests that information provided by cross-stores should be considered.

Apart from that, a subsequent step in preprocessing aimed to identify whether the analyzed week could be considered as a promotional or a regular pricing-week. This categorization was performed for each product, by setting the week as promotional (or regular) when at least 40% of the stores had promotional prices (or regular) prices). Figure 3.5 (a) and Figure 3.5 (b) show that prices are scarce and corrupted by impulsive noise for both promotional and regular prices. To overcome these problems and get the same prices for all stores according to HQ pricing policy, a twofold preprocessing is performed. First, the well-known median filter has been used as the estimation operator, denoted as $\Phi\{P_{i,k}(t)\} = \text{median}_k\{P_{i,k}(t)\}$. This filter is robust with respect to the statistical distribution of the uncertainty term in Eq. (3.3). We empirically chose a size

window of 5 elements, and the filter was applied every week for all the available stores (i.e. one dimensional filtering). Second, the mode of every week was computed in order to have only one price per week. Two examples of the preprocessing results are shown in Fig. 3.5 (c) and (d) for regular and promotional prices, respectively.

3.3.4 Experiments and Results

In this section, a set of experiments for analyzing the suitability of the methodology presented in Fig. 3.3 are first described. Specifically, the quality of models with increasing algorithmic complexity (DEC, TS, nonlinear models), are benchmarked and compared, as well as with different feature spaces. Then, model quality and reliability are proposed and applied to laundry category database.

Data Model Analysis in a Multidimensional Feature Scale

DEC static model. DEC static model is estimated by considering only each price index own-effect after preprocessing. The sold units' estimator requires a smoothing operator (Eq. (3.8)); to do so, we implemented two methods. The average, obtained as a function of the discrete set of prices, and the k -NN estimator, which provides a statistically more effective effect, limiting the impact of outliers. The k -NN method depends on the number of neighbors considered for local-averaged estimation, explored in the range (1,30) and selected the one minimizing MAE (Eq. (2.45)).

Recall that this DEC estimation does not take into account neither simultaneous promotional effects in related products nor temporal structure. For the rest of this section, DEC model based on the k -NN estimator is considered for several reasons: (1) we experimentally checked that it is robust to outliers; and (2) sold units were estimated for the whole range of prices, not only for a discrete set of prices, being able to benchmark results with those provided by other models proposed in this work. Table 3.8 shows the average obtained from bootstrap resampling for merit figures MAE and ΔCI when using DEC k -NN estimator. It can be seen that this method performs similarly in terms of mean and scatter. Results interpretation based on consumer behaviors suggest that, in general, sales estimations for products with a higher number of promotions have worse quality (e.g., Product 1). Note also that good performance is achieved for the private label (Product 10).

DEC approach can be suitable for promotional sales forecasting, though several limitations can be observed. First, the expected effects of demands with respect to prices are not clearly evidenced, even when we use a high rotation category without seasonal effects. Second, according to the observed data, similar prices can yield very different number of sold units, and sometimes lower prices seem to result in lower demand. However, it could be argued that static models for

Table 3.8: Bootstrap test for k -NN DEC. For each cell, the table shows free parameter k in parentheses (first row), average of MAE (second row) and ΔCI (third row) obtained from bootstrap resampling.

Model	$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$
DEC	(1)	(1)	(11)	(7)	(30)	(27)	(24)	(30)	(3)	(13)
	78.38	32.67	20.78	156.80	273.32	34.92	30.87	34.52	29.15	28.27
	187.39	83.47	55.11	675.75	1079.84	103.51	101.20	78.53	90.67	81.01

non-perishable products can obviate temporal conditions. For example, consumers could easily defer laundry detergent purchase, or they could accumulate a number of laundry detergents in a single purchase when prices are low. However, on the whole, these apparently contradictory results raise doubts about the suitability of the DEC promotional model for fitting the products in our database.

Intrinsic and exogenous TS dynamic models. Two different TS promotional models were considered, namely, an AR description (Eq. (3.9)), and an ARx promotional model (Eq. (3.10)). In both cases, we used a two hold-out technique (50% for training) to estimate their out-of-sample performance. We explored orders p and q up to 10 lags, selecting the ones which minimize MAE (Eq. (2.45)).

Table 3.9 shows the p -th and q -th selected orders in terms of MAE for the AR and ARx models, respectively; and the average obtained from bootstrap resampling for merit figures MAE and ΔCI for each product. We obtained non parsimonious models with high orders for both p and q , which highlights a mismatch between the model proposed by TS and the data dynamics. For some products (3, 4, 7, 8, and 9), the time series of the sales volume seemed self-related and with limited correlation with the exogenous variable (prices), whereas for the other products, the performance improved significantly when the exogenous variable was considered. Nonparametric paired bootstrap resampling method was applied to test whether the differences in the benchmarking comparison in the table were statistically significant.

Furthermore, Table 3.9 presents the following comparisons: (1) DEC versus AR; (2) DEC versus ARx; and (3) AR versus ARx. From the first comparison, we can conclude that there were significant performance differences in ΔMAE for most products (except Products 8 and 9), indicating that DEC yielded significantly better quality for the estimations, and the scatter was lower when DEC was considered for Products 2, 6, 7 and 8. The second comparison indicated significant performance differences in ΔMAE for Products 3, 4, 5, 6, 7, 8 and 9, showing that DEC yielded better quality for the estimation, in contrast, for Product 1 ARx yielded significantly better quality. However, the scatter was lower when we used DEC for Product 1, indicating significantly better predictions in terms of scatter. The third comparison indicated significant performance differences were found in ΔMAE for the Products 3, 7, 8 and 9, demonstrating that

Table 3.9: Individual and paired bootstrap tests for TS (AR, ARx). Individual: For each cell, the table shows p -th and q -th selected orders (first row), the average MAE (second row), and 95% CI (third row) from bootstrap resampling. Paired bootstrap between DEC and TS methods, and between AR and ARx: average of ΔMAE (first row) and ΔCI (second row). For each product, boldface emphasize that in the comparison $Model_1$ vs $Model_2$, best performance is achieved with $Model_1$ (negative values) or with $Model_2$ (positive values).

Model		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$
Individual	AR (p)	(10)	(10)	(10)	(9)	(10)	(10)	(10)	(10)	(10)	(9)
		132.26	59.46	30.14	221.99	373.92	72.92	67.85	37.48	20.80	42.16
		156.87	126.48	72.04	1050.68	1376.73	227.44	150.44	85.32	54.01	102.58
	ARx (p, q)	(10,10)	(10,10)	(10,10)	(10,10)	(10,10)	(10,10)	(10,10)	(10,10)	(10,10)	(10,10)
		64.75	33.20	43.25	233.05	313.11	56.87	93.91	97.88	56.04	24.79
	246.07	74.89	172.68	860.55	876.86	183.60	332.07	598.53	296.87	64.77	
Paired	DEC vs AR	-53.72	-26.88	-9.28	-65.54	-98.90	-37.81	-37.01	-2.77	8.24	-13.94
		30.38	-43.00	-17.39	-391.09	-290.65	-121.53	-49.34	-6.85	36.48	-21.20
	DEC vs ARx	13.84	-0.63	-22.24	-77.27	-39.54	-22.01	-63.65	-63.61	-26.80	3.58
		-129.16	33.12	2.98	-60.97	153.75	92.95	2.86	19.01	-42.78	4.97
	AR vs ARx	67.29	26.40	-13.05	-11.66	60.51	15.83	-26.57	-61.73	-35.24	17.43
	-89.21	51.89	-100.50	199.32	499.59	43.64	-176.22	-516.22	-244.36	37.95	

AR methods performed significantly better for the estimation, whereas for Products 1, 2, 6 and 10 ARx yielded a better quality. The scatter was lower when AR was considered for Products 2, 5 and 10, yielding significantly better predictions in terms of scatter, and ARx for Products 1, 3, 7, 8 and 9. In summary, no clear trend in terms of general behavior and prediction of the promotional models could be observed in this set of models.

Improvements from nonlinear methods. k -NN method for nonlinear promotional modeling was used. Its design depends on a free parameter, k , which stands for the number of neighbors considered for local-averaged estimation. In this study, the range (1,30) was explored and selected the value which provided the minimum MAE.

Different feature vectors to characterize temporal evolution in terms of exogenous and/or endogenous variables were explored. The notation for the feature space in this experiment, for a temporal depth n_0 , in terms of the feature space, is as follows:

$$\Xi^{t_0} = [\hat{P}(t), \dots, \hat{P}(t - t_0)] \quad (3.14)$$

$$\Theta^{t_0} = [S(t - 1), \dots, S(t - t_0)] \quad (3.15)$$

$$\Psi^{t_0} = [\Xi^{t_0}, \Theta^{t_0}] \quad (3.16)$$

According to Eq. (3.14), which addresses different temporal depths for past prices, five models were benchmarked, i.e. $\Xi^1, \Xi^2, \Xi^3, \Xi^4, \Xi^5$. Results showed that the estimated performance

Table 3.10: Individual and paired bootstrap tests for k -NN. Individual: For each cell, the free parameter k (first row), the average MAE (second row), and 95% CI (third row) from bootstrap resampling. Paired bootstrap between k -NN and DEC, k -NN and TS, and k -NN with different temporal depth: average of ΔMAE (first row) and ΔCI (second row). For each product, boldface emphasize that in the comparison $Model_1$ vs $Model_2$, best performance is achieved with $Model_1$ (negative values) or with $Model_2$ (positive values).

Model		$i = 1$	$i = 2$	$i = 3$	$i = 4$	$i = 5$	$i = 6$	$i = 7$	$i = 8$	$i = 9$	$i = 10$
Individual	k -NN $_{\Xi^1}$ (k)	(2)	(30)	(9)	(1)	(30)	(30)	(22)	(30)	(2)	(20)
		75.25	33.74	21.13	88.89	292.55	35.85	30.10	33.60	28.92	25.25
		170.12	87.95	56.96	350.75	1251.68	135.27	95.47	74.84	87.22	61.13
	k -NN $_{\Psi^1}$ (k)	(2)	(2)	(3)	(8)	(17)	(5)	(1)	(3)	(2)	(20)
		67.42	47.67	21.70	208.75	268.92	67.68	38.00	39.57	33.93	28.67
		141.17	107.98	53.45	580.44	1053.12	152.35	85.50	93.84	73.47	73.87
Paired	DEC vs k -NN $_{\Xi^1}$	6.77	-2.35	0.07	4.06	23.76	-7.08	2.20	0.90	0.00	-0.69
		14.11	-8.12	1.83	-17.45	36.80	-57.74	8.51	-10.43	-4.25	7.73
	AR vs k -NN $_{\Xi^1}$	57.62	26.10	9.12	133.21	83.07	36.44	38.01	3.63	-8.26	16.56
		-15.14	45.81	15.67	700.05	123.79	105.61	54.34	11.10	-33.67	40.72
	ARx vs k -NN $_{\Psi^1}$	-3.98	-14.72	21.71	24.16	47.60	-10.78	56.31	59.47	22.64	-3.82
		103.81	-24.77	118.69	260.35	-166.01	44.98	240.22	501.94	225.12	-8.58
Paired	k -NN $_{\Xi^1}$ vs k -NN $_{\Psi^1}$	8.08	-14.13	-0.59	-120.94	24.07	-32.01	-7.66	-5.92	-4.89	-3.54
		28.57	-19.67	3.33	-237.22	200.80	-18.69	10.58	-19.08	13.96	-13.30
	k -NN $_{\Psi^1}$ vs k -NN $_{\Psi^2}$	3.71	0.77	-0.00	0.64	-11.29	-0.52	-0.38	0.19	-0.31	-0.59
		13.75	-0.82	0.22	-11.95	-40.21	-0.71	-17.87	0.18	0.13	-7.22

improved when including in the model information over two consecutive weeks, current and past. Thus, Ξ^1 was a suitable feature space for nonlinear promotional models, when considering only the exogenous variable. A similar analysis using a set of consecutive temporal depths n_0 for exogenous and endogenous variables simultaneously was performed, observing better results when considering information over coupled consecutive weeks, this is, for Ψ^1 .

Table 3.10 shows the individual bootstrap tests in terms of MAE and CI when considering Ξ^1 and Ψ^1 since both are the feature spaces that provide the minimum error.

Table 3.10 also shows the results when comparing static, dynamic, linear and nonlinear models with paired bootstrap tests. First, regarding static and dynamic models (DEC vs k -NN $_{\Xi^1}$), significant performance differences in ΔMAE were observed for Products 1 and 7, for which the dynamic approach yielded better estimation quality, whereas for Product 6, the static model performed better. Secondly, we compared linear and nonlinear models, namely, AR vs Ξ^1 , obtaining better quality in terms of scatter for Products 2, 4, 6, 7 and 10 when the nonlinear model is considered. Regarding linear and nonlinear models with past sold units and prices, it

is observed significantly better estimation quality in ΔMAE for Products 3, 7, 8 and 9, whereas nonlinear models showed better results for Product 2. It is noteworthy that the scatter was lower when nonlinear models were considered for Products 1, 3, 7, 8 and 9.

Thus, it can conclude that, for modeling the sold units of laundry detergent products in this database, the consideration of modeling promotional sales with nonlinear methods yields a significant performance improvement for most of the products. Therefore, we benchmarked new feature vectors with different temporal depths, which indicated that including additional past sold units and prices did not improve sales forecast. Table 3.10 also shows that, for Products 2, 4, 6, 7, 8 and 9, it was significantly better to exclude the endogenous variable (sales).

Model Quality and Reliability Indicators

In contrast to conventional approaches, which use a single value of a merit figure to evaluate model performance, it is proposed in this Thesis the use of a new set of indicators to characterize both quality and reliability in a given region $R \in \Psi_i$ (Ψ_i^R). Note that different values for the same indicator can be obtained for different regions. As an intuitive example, indicators may be less reliable in regions with scarce, noisy or non-informative data.

A set of four quality indicators will be calculated using the B estimations $u_i^*(b) = F(V^*(b))$, obtained through the B resamples $\{V^*(b)\}_{b=1}^B$. These indicators are proposed to measure the reliability of model F , and are defined as follows.

(1) *Variation Coefficient (VC) Index.* VC measures dispersion in relation to mean value. It is a useful statistic for comparing the degree of variation between two datasets, even when their means are drastically different. The lower the VC , the more reliable our predictions are. It can be written as

$$VC_{\hat{S}_i}^R(\Psi_i) = \frac{\sigma_{\hat{S}_i}^R(\Psi_i)}{\mu_{\hat{S}_i}^R(\Psi_i)} \quad (3.17)$$

where $\sigma_{\hat{S}_i}^R(\Psi_i)$ and $\mu_{\hat{S}_i}^R(\Psi_i)$ are the standard deviation and mean, respectively, of sales estimations in region R .

(2) *Confidence Intervals Variation (ΔCI).* We particularize ΔCI previously defined in Eq. (2.57) to calculate the reliability of the estimated sales in region R for the i -th product. The narrower the confidence interval, the lower the variability is, hence ΔCI can be used as a reliability measurement in a region R of the feature space, denoted as

$$\Delta CI_{\hat{S}_i}^R(\Psi_i) = CI_{\hat{S}_i}^{R,u}(\Psi_i) - CI_{\hat{S}_i}^{R,l}(\Psi_i) \quad (3.18)$$

where $CI_{\hat{S}_i}^{R,u}(\Psi_i)$ and $CI_{\hat{S}_i}^{R,l}(\Psi_i)$ denote the upper and lower limits of the confidence interval, respectively.

(3) *Baseline Relative Index (BLRI).* Marketing managers widely use baseline sales to assess the profitability and effectiveness of marketing activities by investigating how promotions can

affect baseline sales over time. In this setting, it is necessary the creation of a new index to assess the accuracy of a promotional model in not only absolute but also in relative terms respect to the baseline. This can be achieved by normalizing the number of estimated sold units with respect to the baseline sales, and we define it as

$$BLRI_i^R(\Psi_i) = \frac{\hat{S}_i^R(\Psi_i) - BL_i^R(\Psi_i)}{BL_i^R(\Psi_i)} \quad (3.19)$$

where $BL_i^R(\Psi_i)$ is the estimated baseline sales in region R of the feature space for the i -th product. Note that $BLRI = 0$ indicates that the estimated number of sold units is similar to baseline, whereas $BLRI > 1$ indicates that estimated sold units are greater than baseline.

(4) *Dynamic Range Index (DRI)*. It is based on dynamic range DR_i , defined as the difference between the maximum and minimum values of a variable, for us, estimated sales for the i -th product. We define it as follows:

$$DRI_i^R(\Psi_i) = \frac{\hat{S}_i^R(\Psi_i) - DR_i^R(\Psi_i)}{DR_i^R(\Psi_i)} \quad (3.20)$$

This way, DRI provides an idea of the accuracy in terms of the forecasting variability. The greater the DRI , the lower the variability is.

The four previous indices allow us to check for the reliability and uncertainty of a given model for promotional sales depending on the feature space. Note that the two first (two last) indices are absolute (relative) magnitudes, and that the statistical distribution of \hat{S}_i has to be estimated in the feature space.

As described in Sec. 3.3.4, the statistical distribution of estimated sold units \hat{S} for the i -th product in the space defined by feature input vector Ψ , i.e., $p_{\hat{S}(\Psi)}$, can be readily estimated by using bootstrap resampling, and it is denoted as $p_{\hat{S}(\Psi)}^*$. Its statistical average is a hypersurface of the sales as a function of the feature space, and more general, it provides useful information for both reliability and decision-making point of views by allowing us to obtain the indices previously defined in Eq. (3.17), (3.18), (3.19), and (3.20).

For this database, it was checked that better results were obtained when sold units predictions were made with a nonlinear model considering two consecutive weeks. With this forecasting model, we checked the reliability and stability of results when working with one or two years. This experiment presents the proposed indices in the feature space for two illustrative example products, namely, Product 1 and Product 5. Figure 3.6 shows the predicted sales units \hat{S}_i as a function of the feature space for both products, when using two years (a,c) and one year (b,d). As previously described, changes in the dynamics and the promotions in the available time periods were determinant for the model forecasting capabilities. VC was larger in general for Product 1 with two years data, but also it was larger, in general, in promotional regions of the feature space (e,f). For Product 5, VC was quite constant throughout the feature space, but lower when only

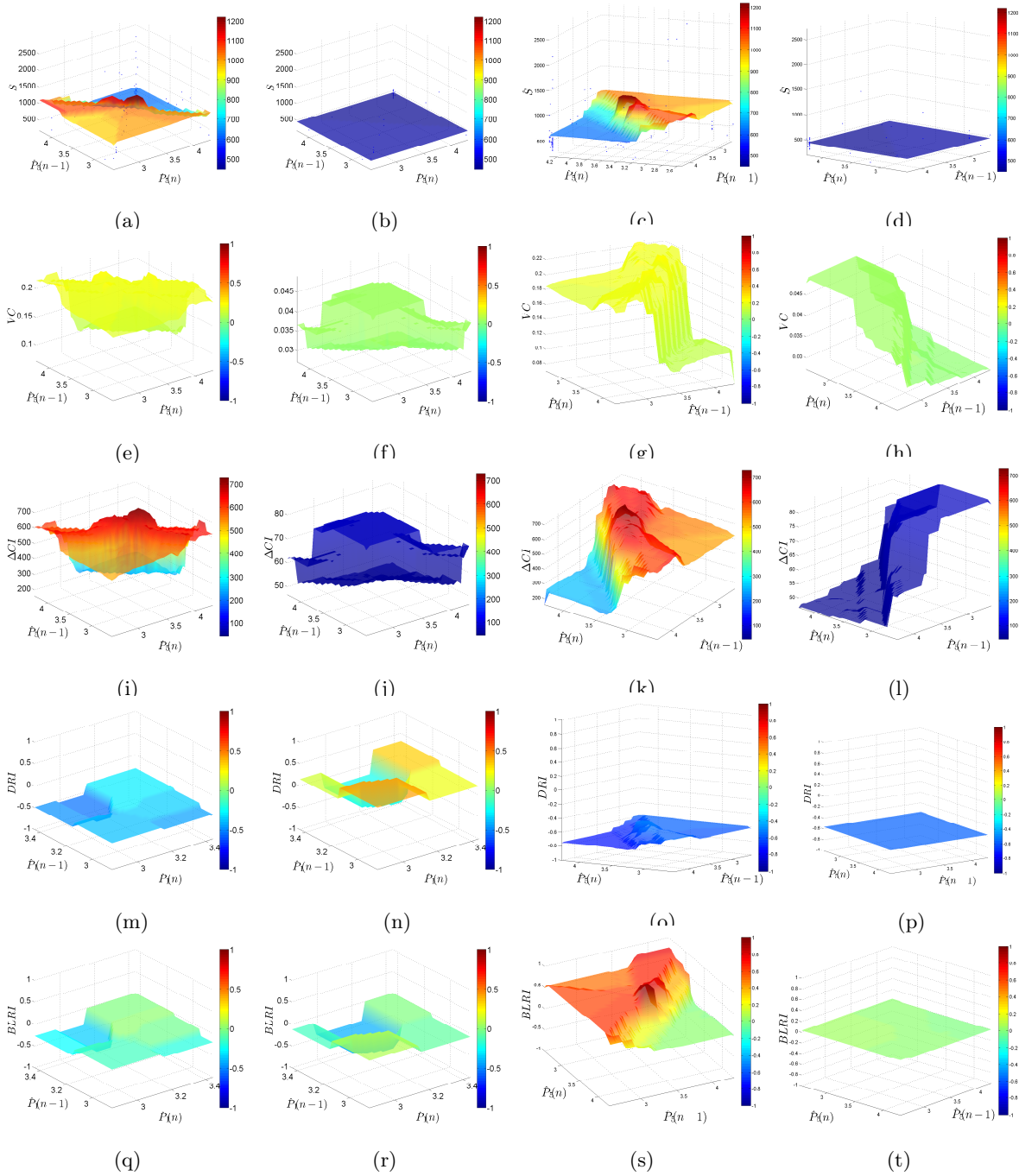


Figure 3.6: Estimated sold units and reliability indices. Columns 1 (two years) and 2 (one year) for Product 1; Columns 3 (two years) and 4 (one year) for Product 5. Estimated sold units: (a) - (d); VC : (e) - (h); ΔCI : (i) - (l); $BLRI$: (m) - (p); DRI : (q) - (t).

considering the last year. ΔCI was strongly dependent of the region in the feature space (i,k), and higher for promotional regions. $BLRI$ for Product 1 indicated a higher efficiency, relative

to the baseline, of the promotional activity when data from one year was used (m,n). *DRI* was larger in promotional regions for Product 5 using the data from two years, whereas it was reduced and became near constant for the data from one year (s,t).

The significant differences among estimated sold units and reliability indices in one year versus two years, indicating that memory effect may not be justified to be larger than one year, and so, proposed models would not necessary provide stability and reliability capabilities, adding information beyond one year historical data.

It is also remarkable the higher stability and more solid behavior of Product 5 in all analyzed statistics over a wider range of different prices. This situation contrasted against Product 1, indicating that for this kind of product significant effects are led by price change.

3.3.5 Discussion and Conclusions

A promotional chain-level analysis and data modeling based on retail aggregated data to support retail marketing decisions is proposed. Figure 3.3 depicts the necessary steps for the reliability and stability analysis of promotional models from a chain-level point of view. First, retail data were aggregated at the store-level using simple preprocessing tools, to come to a market-level decision. Second, linear and nonlinear prediction engines using nonparametric bootstrap resampling based on performance statistics were benchmarked and optimized. Third, we took into account the reliability and stability of the promotional models built with the products in the database using a set of new indicators based on bias and scatter measurements in the feature space.

The economic downturn that began in 2008 is one of the largest in history, at least in some countries; thus, it is more necessary than ever for retailers to effectively evaluate short to medium term promotional effects. It is possible that traditional promotional models do not accurately reflect the actual complexity in the real time because of the increasing amount of concurrent aspects that affect consumer behavior. Therefore, researchers should focus on new models that can capture and statistically represent this new scenario.

Limitations of DEC Static Analysis. Our experiments showed that the DEC analysis could not provide consistent results in terms of unequivocal demand for a certain pricing level in our data. Note that laundry detergent category is not subjected to short expiration date, thus, it could be argued that households may stockpile the product if prices justify doing so. This assumption is one of several behind the “buy two and get one free” promotional offers, which are common for many long-expiration date products. The present study shows that the static DEC model does not provide a direct statistical match between the endogenous (demand) and exogenous (price) variables.

Limitations of TS Linear Models. The existence of communication networks and consumer

networking may change the market dynamics, resulting in a large and increasing number of concurrent effects. Accordingly, retailers have used dynamic TS models based on historical and current data to guide the promotional strategies. In our data, TS linear models with intrinsic and exogenous variables had an acceptable fit to the data, but only at the expense of non-parsimonious models. Although previous research shows that future demand is forecasted better by considering memory for both pricing and demand data, paradoxically, our data set contained several products for which the incorporation of the exogenous variable into the AR analysis significantly worsened forecasting performance.

Scope of Nonlinear Models. In general, non-linear models' predictive capability was more effective than that of linear models. Although the results were not uniform for all the products, we obtained consistently parsimonious promotional nonlinear models that yielded acceptable forecasting performance in all the products with up to one lag for both the exogenous and endogenous variables. In some cases, TS linear models performed similarly to nonlinear ones, which is consistent with the fact that linear TS models are specific cases of nonlinear models.

We observed most behavioral singularities in the premium product (higher price, Product 1) and in the most competitive products (private label and low prices, Products 7, 8, 9, and 10).

Machine Learning for Healthcare Analytics

4.1 Introduction

Electronic Health Records (EHRs) are collections of health information in digital storage format, which can in theory be shared among systems to convey the relevant information of a patient [26]. EHRs have three levels of medical understanding, namely, data storage, information, and knowledge [122]. While technology seems to have successfully covered the data storage level, the others are currently intensive research tasks. In the last decades, a considerable amount of literature exists on knowledge extraction from the EHRs, aimed to support clinical decision-making in several domains [27, 28, 29, 30, 31, 32, 33]. In this chapter, EHR data related to the gastrointestinal surgery domain are analyzed to address different goals: (1) to detect complications after CRC at an early stage; and (2) to predict surgical site infections (SSI) at both pre-operative and post-operative stages for patients admitted for gastrointestinal surgery.

According to American Cancer Society, CRC is the third most common cancer diagnosed in both men and women in developed countries, being the surgery the only curative treatment [34]. Nevertheless, the elective colorectal resection is normally associated with a complication rate of 20-30% [35], being reported that AL occurs in 5-15% of all patients who underwent CRC surgery [36]. Early diagnosis and intervention can minimize systemic complications, and can be vital in the case of AL due to it may be a lethal condition. However, it is hindered by current diagnostic methods that are non-specific and often uninformative [37], thus, novel methods are required to identify and detect this complication at an early stage using EHR data.

On the other hand, SSIs are among the most common hospital-acquired infections. In fact, they represent up to 30% of all hospital acquired infections [123, 124]. SSIs are associated with considerable morbidity and mortality. A mortality rate of 3%, prolonged stay up to 10 days and a significant decrease in quality of life, are reported. Similarly, readmissions related to SSIs are associated with a considerable increase in healthcare cost, up to 27,000 USD per

readmission [125]. This persistent in-hospital morbidity is particularly associated with surgery for CRC [126, 127, 128].

In this Thesis, we first focus on the task of early detection of AL using free text extracted from EHRs. Then, we propose a learning system architecture capable of jointly exploiting heterogeneous sources for AL early detection. We used linear and non-linear kernel methods individually for each data source, and leveraging the powerful composite kernels for their effective combination. Finally, we built a prediction model for pre-operative and post-operative SSI using different methods to manage the sparsity of the EHR data sources.

4.2 Application 4.1: Early Detection of Anastomosis Leakage from Bag-of-Words

4.2.1 Introduction

A considerable amount of literature exists on extraction of knowledge from EHRs to support clinical decision-making (see [30] and references therein). Specifically, analysis of the (unstructured) EHR free text may potentially extract a large amounts of information regarding patient health status and medical history, which may not be fully available in the structured data that are also available in EHRs [31, 32].

ML methods have recently demonstrated great potential at free text analysis for decision support and medical information retrieval. Several such methods are based on the simple, but often powerful, Bag-of-Words (BoW) model. Wright et al. [32] used this model to identify relevant documents in EHRs pertaining to a user's query on progress notes in diabetes, and in [129], a system for automatic case identification was proposed for observational epidemiological studies. Using various levels of sophistication in the BoW model, the authors in [31] developed a framework for general-purpose automatic diagnosis in traditional Chinese medicine. Furthermore, the authors in [130] derived a semi-supervised SVM, for automated identification of primary care records from the General Practice Research Database, with applications to retrieval of coronary angiogram and ovarian cancer diagnoses, and in [131] a comprehensive bag-of-concepts system was proposed for quantifying a patient's risk of mortality and complications. The interested reader can also see reference [132] for a recent review of natural language processing techniques for analysis of free text in EHRs, in addition to [133] for a review on extracting information from textual documents in EHRs, including the advances in the field from 1995 to 2008. However, few studies have explored systematic *FS criteria* for ML based applications using EHR data, or principled knowledge extraction from the ML engines.

In this Thesis, the detection of AL using a BoW model extracted from an EHR is analyzed. This work was based upon a patient database (QUAKE, quality control of surgical performance

4.2 Application 4.1: Early Detection of Anastomosis Leakage from Bag-of-Words85

with unstructured EHR data) which was extracted from the Department of Gastrointestinal Surgery at the University Hospital of North Norway. First, different ML techniques are benchmarked using pre-operative and post-operative data. Then, the early detection of AL was further explored. Several novel FS strategies described in detail in Sec. 2.4.2, that are capable of automatically identifying the relevant words, while permitting easy knowledge extraction from the system, are applied.

A vast general literature exists on FS, see Sec. 2.4 for examples. As novel alternatives to the RFE, innovative FS methodology in order to avoid numerous SVM re-training procedures is proposed in Sec. 2.4. Our present work introduces statistically principled FS methods, capable of working on the linear classifier weight amplitudes in an easy way with extremely high dimensional input spaces. The proposed methods require no pre-specification of the number of features to obtain, and are based on three different criteria (see Sec. 2.4.2 for details).

After adjusting for imbalanced classes, which is a well-known challenge in medical classification applications [129, 134], the proposed FS strategies are shown to significantly improve the detection of AL. Also, the results provide useful knowledge of the relevant words (without need of their pre-selection by clinicians) and their temporal evolution.

4.2.2 Database

The database used in the current study consisted of unstructured Norwegian text extracted from the EHR used at the Department of Gastrointestinal Surgery at the University Hospital of North Norway. All documents related to both inpatient and outpatient visits between 2004-2012 were extracted. The most frequent document types that were extracted were nurses' notes, journal notes, outpatient notes, radiology reports, referrals, discharge letters and admission notes. A clinician manually reviewed the EHR of 402 patients admitted for CRC surgery in 2006-2011, and 31 patients with AL were identified. The negative class consisted of the 371 remaining patients.

A BoW model was subsequently built, by counting all unique words appearing in the database. There were a total of 65328 unique words in the database. Hence, the database is represented as $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$ where each \mathbf{x}_i , representing the i -th patient, is 65328-dimensional. For compact notation, we collect the data samples \mathbf{x}_i , $i = 1, \dots, n$, in the matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$. In the linear SVM $y = \langle \mathbf{w}, \mathbf{x} \rangle + b$, each element, or feature, in \mathbf{x} hence corresponded to the number of appearances in a EHR for a given patient of one of the unique words.

Preprocessing. Initially all words were transformed to lowercase and all grammatical symbols were removed. Furthermore, all numbers and stop words were filtered out. Apart from that, advanced natural language processing procedures, such as combining words with identical meanings or corrections of obvious misspellings, were not considered in this work.

These unique words represent the "bag" in the BoW model. The bag cardinality was reduced by keeping only those words appearing at least a certain number of times (assuming that, for instance, misspelled words appear relatively infrequently). In this work, we empirically kept only words appearing at least 10 times, reducing the dimensionality of the vectors \mathbf{x}_i from 65328 to 13188. Of course, enforcing a threshold may lead to information loss. Note that enforcing a too high threshold may lead to information loss. In the remainder of this work, the data set consists of the resulting 13188 words.

In previous general-purpose text classification studies using SVM [135, 136, 137, 138], normalization has been suggested for preprocessing. Normalization may be obtained in several ways. Term frequency - inverse document frequency (TF-IDF) representation [139] is a common method. Here this and other normalization strategies, such as standardization to mean zero and unit variance, were considered. Alternatively, feature vectors may be normalized to equal (Euclidean) length. In this study, such normalizations did not influence the results much, and they were not pursued further.

Finally, the feature set can be represented on a binary basis, by the presence or absence of each word, so that the influence of high frequency words that do not necessarily exhibit discriminatory power is reduced. This binary dataset is denoted by \mathbf{X}^{bin} .

4.2.3 Experiments and Results

Experimental Setup

This experiments section starts by analyzing and discussing the tuning of the free parameter ν in the SVM, and then comparing the SVM AL classification performance on the dataset \mathbf{X} without FS, with those of FDA and NB (see Sec. 2.3.1 for more details). We subsequently analyze in detail the effect of the proposed FS strategies, and show that results improve significantly. Finally, a temporal analysis explores the viability of early detection of AL by means of the BoW model.

Parameter Tuning. The linear ν -SVM algorithm requires the tuning of a single free parameter $\nu \in (0, 1)$, which has to be tuned. This parameter must be tuned based on the available training set. We adopted a LOO strategy for the tuning of ν , ensuring that the parameter tuning was always based on out-of-sample performance. For completeness, we evaluated several different performance measures, namely, Pe , Se , Sp , and BER (see Sec. 2.5.1).

Classification problems are frequently imbalanced. For example, in the binary case, the number of samples in the positive class may be substantially smaller than the number of samples in the negative class. Several previous ML studies have shown that balanced classes in the training data set provide improved overall classification performances (see e.g. [140] and references therein). Common strategies to balance the classes include undersampling, i.e., removing samples

4.2 Application 4.1: Early Detection of Anastomosis Leakage from Bag-of-Words87

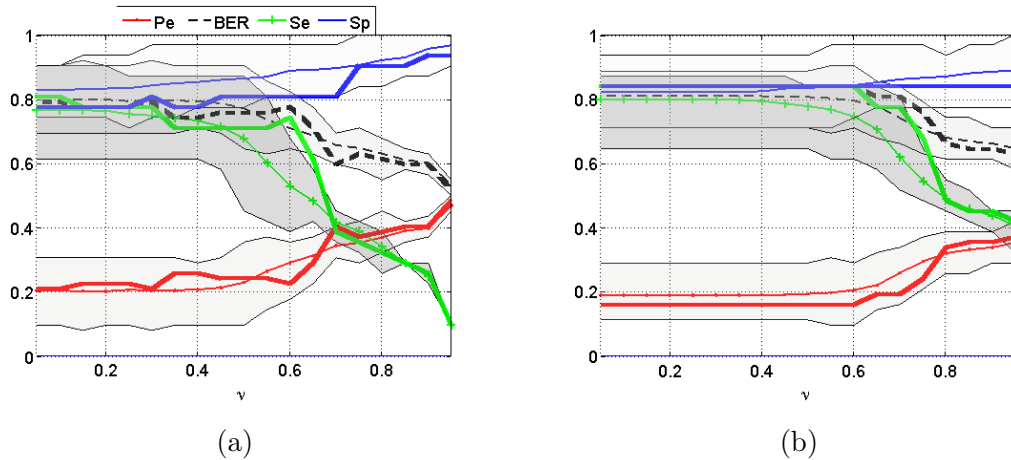


Figure 4.1: Free parameter (ν) tuning in terms of several figures of merit (Pe , Se , Sp and BER) for random (thick) and resampling (fine and filled, CI shaded) downsampling, evaluated for \mathbf{X} and \mathbf{X}^{bin} in (a) and (b), respectively.

from the majority class, at the risk of information loss, or oversampling the minority class has also been studied, at the risk of overfitting.

The training set was constructed using an undersampling strategy in order to enforce balanced classes. Towards that end, a *random subset* (31 samples) of the negative class was selected, together with the 31 positive samples in the database. This random subset was used for the tuning of ν . The results, one for each performance measure, are shown in Fig. 4.1, indicated by the thick line (see figure text for further explanation). Observe that the best performance was obtained for a relatively wide range of smaller values of ν , independently of the figure of merit used. CM was computed for $\nu \in [0.05, 0.4]$ (not shown) finding that the error rates were basically the same over this range of ν . In the end, a value of $\nu = 0.05$ was used in subsequent experiments (see below). The reason for this choice was that ν represents an upper bound on the fraction of margin errors and a lower bound of the fraction of support vectors relative to the total number of training examples. As few support vectors as possible, while maintaining performance, is in general considered a positive property of any SVM method.

In order to analyze the appropriateness of the particular random subset used here, in a statistical sense, we extracted further 50 random resamples (with replacement) from the negative class. Figure 4.1 shows the mean performance (fine line, see figure text) and the CI (filled tube) for each of the figures of merit. It is important to note that the results corresponding to the initial random sample lie well within the CI, and may therefore be considered representative for the negative class.

The test or generalization performance of the SVM received special attention in this work. The key element when evaluating the generalization ability is to keep the training and the testing

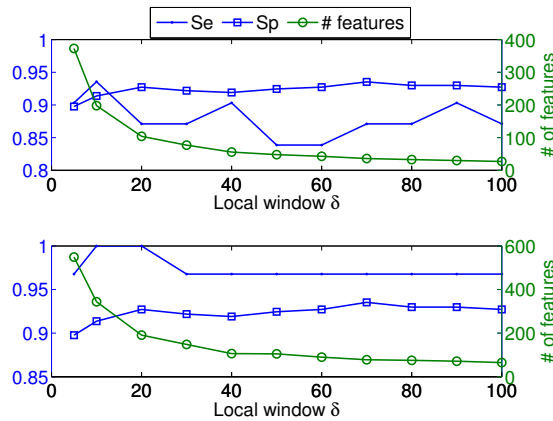


Figure 4.2: Number of features, Se and Sp depend on the size of a local window δ of neighbor weights for \mathbf{X} (upper panel) and \mathbf{X}^{bin} (lower panel).

process independent, as far as possible in the given database. For this purpose, the overall data set was divided in two parts, one part in which there is a balance between positive and negative instances (balanced part), and a second part consisting of the remaining negative instances. The generalization ability was measured by a two-stage process. The first stage invokes a LOO cross-validation scheme on the balanced part. In the second stage, a SVM classifier is constructed using the whole balanced set and it is used to classify the remaining negative instances. Results in two stages are combined. Table 4.1 shows CM for the SVM, together with the FDA and NB methods (using the two-stage process), for both feature spaces \mathbf{X} and \mathbf{X}^{bin} (SVM only). First of all, the table shows that FDA and NB performances are clearly lower to those of the SVM. Interestingly, for the SVM, results were better on \mathbf{X}^{bin} compared to \mathbf{X} . We used a paired bootstrap resampling test as proposed in Sec. 2.5.2 to establish statistical significance of the different performances across methods, obtaining that \mathbf{X}^{bin} performs better than the other ones.

When using FDA, it is well-known that the inherent matrix inversion is problematic when the number of features, i.e. the dimensionality, is greater than the number of samples. For that reason, the dimensionality of the feature vectors was forced to be less than 402, which was the number of samples, by considering the 350 most frequently occurring words. A problem when using NB, is that some of the most infrequent words, or features, are not appearing in both classes. In order to avoid this problem, only those features appearing in both classes were kept.

In the testing phase, the classes were imbalanced. For this reason, we also display the performance, or generalization ability, of the SVM in terms of Se and Sp in Table 4.1. The SVM results on \mathbf{X}^{bin} also stand out with respect to Se measures.

4.2 Application 4.1: Early Detection of Anastomosis Leakage from Bag-of-Words89

Table 4.1: Performance for ν -SVM, FDA, and NB classifiers.

	ν -SVM, \mathbf{X}	ν -SVM, \mathbf{X}^{bin}	FDA, \mathbf{X}	NB, \mathbf{X}
CM	$\begin{bmatrix} 25 & 56 \\ 6 & 315 \end{bmatrix}$	$\begin{bmatrix} 26 & 52 \\ 5 & 319 \end{bmatrix}$	$\begin{bmatrix} 15 & 208 \\ 16 & 163 \end{bmatrix}$	$\begin{bmatrix} 10 & 28 \\ 21 & 343 \end{bmatrix}$
Se	81%	84%	48%	32%
Sp	85%	86%	42%	92%

Feature Selection

In this section, the attention is turned to the analysis of the proposed FS strategies, namely, a simple statistical criterion (LOO based test), an intensive-computation statistical criterion (bootstrap resampling), and an advanced statistical knowledge criterion (kernel entropy), see Sec. 2.4.2. The core idea is the following: 1) a subset of relevant features was selected by one of the proposed algorithms, and 2) the linear ν -SVM classifier was retrained with the selected features and used to classify test samples. As shown below, the performance of the classifier *increases* as a result of the FS.

First, a brief discussion on the selection of free parameters in the FS algorithms is provided. For bootstrap resampling, the free parameter corresponds to the size of a local window δ of neighbor weights. Small δ values provide higher number of selected features, whereas the opposite is true for larger δ values. This is illustrated in Fig. 4.2 where Se and Sp results are shown based on bootstrap FS retraining over a range of δ values. Results suggest that good performance was obtained when considering $\delta = 10$, which is the value used in subsequent experiments.

Kernel entropy component FS requires the selection of the tail probability and the kernel size (σ value). We have experienced that a tail probability of 0.05 provides good results. Furthermore, since w is a one dimensional random variable, Silverman's rule [141] for kernel size selection is known to be reliable, and for that reason that criterion was used in the remainder. With this approach, kernel size is obtained as follows: $\sigma = 1.06std(\mathbf{w})N^{-1/5}$, where std is the standard deviation and \mathbf{w} is the weight vector obtained for each dataset.

Table 4.2 shows the benefit of FS in terms of CM , Se , Sp and the number of selected features obtained by the LOO based test, the bootstrap resampling, and the kernel entropy criterion for both databases \mathbf{X} and \mathbf{X}^{bin} . The power of the proposed FS methods can be observed by noting that all of them *improve* Se and Sp measures. Furthermore, these improvements were obtained by using *far fewer features*, compared to the original dimensionality of the data.

Results suggest that the best performance is obtained with the bootstrap resampling approach for both \mathbf{X} (Se 100%, Sp 89%) and \mathbf{X}^{bin} (Se 100%, Sp 89%). The number of features selected to obtain these results were 196 for \mathbf{X} and 292 for \mathbf{X}^{bin} .

For completeness, the proposed FS strategies were compared with the RFE method [68].

Table 4.2: *FS criteria analysis. CM, Se, Sp and number of selected features obtained by LOO based test, bootstrap resampling, and kernel entropy criterion (Keca) for \mathbf{X} (upper) and \mathbf{X}^{bin} (lower) inputs spaces.*

	All	LOO	Boot ($\delta = 10$)	Keca
CM	$\begin{bmatrix} 25 & 56 \\ 6 & 315 \end{bmatrix}$	$\begin{bmatrix} 28 & 52 \\ 3 & 319 \end{bmatrix}$	$\begin{bmatrix} 31 & 39 \\ 0 & 332 \end{bmatrix}$	$\begin{bmatrix} 25 & 55 \\ 6 & 316 \end{bmatrix}$
Se	81%	90%	100%	81%
Sp	85%	86%	89%	85%
# features	13188	6896	196	212
CM	$\begin{bmatrix} 26 & 52 \\ 5 & 319 \end{bmatrix}$	$\begin{bmatrix} 29 & 52 \\ 2 & 319 \end{bmatrix}$	$\begin{bmatrix} 31 & 42 \\ 0 & 329 \end{bmatrix}$	$\begin{bmatrix} 31 & 45 \\ 0 & 326 \end{bmatrix}$
Se	84%	94%	100%	100%
Sp	86%	86%	89%	88%
# features	13188	8073	292	189

Table 4.3: *Proposed FS benchmarked with RFE for non-binary (\mathbf{X} , upper) and binary (\mathbf{X}^{bin} , lower) input feature spaces.*

	LOO	RFE	Boot	RFE	Keca	RFE
Se	90%	87%	100%	87%	81%	80%
Sp	86 %	86%	89%	82%	85%	82%
# features	6896	6896	196	196	212	212
Se	94%	90%	100%	100%	100%	100%
Sp	86%	86%	89%	88%	88%	88%
# features	8073	8073	292	292	189	189

Results obtained using the proposed FS methods and RFE are shown in Table 4.3, using for clarity the same number of features for RFE as the number of features selected by the proposed methods, respectively. Recall that RFE requires the training of multiple classifiers on subsets of features of decreasing size, and for this reason, it does not trivially provide the optimum number of features to be selected. We also implemented the RFE cross-validation procedure (requiring up to 12 hours run-time on a standard research-purpose laptop for one data set) obtaining results which were very similar to those displayed in Table 4.3. This shows that the proposed FS methods may extract useful information from the EHRs, similarly or better when compared to the RFE, however, it is based on statistical criteria requiring no pre-specification of the number of features to be selected, nor any computationally demanding cross-validation.

Early AL Detection Experiments

The *early detection* of AL was further explored. Towards that end, several additional databases were created. The databases \mathbf{X}_{op} and \mathbf{X}_{op}^{bin} represented the BoW based on all journal

4.2 Application 4.1: Early Detection of Anastomosis Leakage from Bag-of-Words91

Table 4.4: Temporal analysis (CM and number of features) for different data time slots: up to and including day of surgery; four days after surgery or until patients leave the hospital, for non-binary and binary input feature spaces.

FS	\mathbf{X}_{op}	\mathbf{X}_{+4}	\mathbf{X}	\mathbf{X}_{op}^{bin}	\mathbf{X}_{+4}^{bin}	\mathbf{X}^{bin}
All	$\begin{bmatrix} 19 & 186 \\ 12 & 185 \end{bmatrix}$	$\begin{bmatrix} 17 & 126 \\ 14 & 245 \end{bmatrix}$	$\begin{bmatrix} 25 & 56 \\ 6 & 315 \end{bmatrix}$	$\begin{bmatrix} 20 & 145 \\ 11 & 226 \end{bmatrix}$	$\begin{bmatrix} 22 & 112 \\ 9 & 259 \end{bmatrix}$	$\begin{bmatrix} 26 & 52 \\ 5 & 319 \end{bmatrix}$
Se	61%	55%	81%	65%	71%	84%
Sp	50%	66%	85%	61%	70%	86%
# features	5409	6858	13188	5409	6858	13188
LOO	$\begin{bmatrix} 28 & 193 \\ 3 & 178 \end{bmatrix}$	$\begin{bmatrix} 25 & 118 \\ 6 & 253 \end{bmatrix}$	$\begin{bmatrix} 28 & 52 \\ 3 & 319 \end{bmatrix}$	$\begin{bmatrix} 29 & 131 \\ 2 & 240 \end{bmatrix}$	$\begin{bmatrix} 30 & 93 \\ 1 & 278 \end{bmatrix}$	$\begin{bmatrix} 29 & 52 \\ 2 & 319 \end{bmatrix}$
Se	90%	81%	90%	94%	97%	94%
Sp	48%	68%	86%	65%	75%	86%
# features	2840	3912	6896	2991	3992	8073
Boot	$\begin{bmatrix} 30 & 196 \\ 1 & 175 \end{bmatrix}$	$\begin{bmatrix} 30 & 130 \\ 1 & 241 \end{bmatrix}$	$\begin{bmatrix} 31 & 39 \\ 0 & 332 \end{bmatrix}$	$\begin{bmatrix} 31 & 105 \\ 0 & 266 \end{bmatrix}$	$\begin{bmatrix} 31 & 82 \\ 0 & 289 \end{bmatrix}$	$\begin{bmatrix} 31 & 42 \\ 0 & 329 \end{bmatrix}$
Se	97%	97%	100%	100%	100%	100%
Sp	47%	65%	89%	72%	78%	89%
# features	107	102	196	120	142	292
Keca (5%)	$\begin{bmatrix} 29 & 181 \\ 2 & 190 \end{bmatrix}$	$\begin{bmatrix} 30 & 146 \\ 1 & 225 \end{bmatrix}$	$\begin{bmatrix} 25 & 55 \\ 6 & 316 \end{bmatrix}$	$\begin{bmatrix} 29 & 125 \\ 2 & 246 \end{bmatrix}$	$\begin{bmatrix} 31 & 85 \\ 0 & 286 \end{bmatrix}$	$\begin{bmatrix} 31 & 45 \\ 0 & 326 \end{bmatrix}$
Se	94%	97%	81%	94%	100%	100%
Sp	51%	61%	85%	66%	77%	88%
# features	90	110	212	86	106	189

notes up to and including the day of surgery. At this point in time, none of the patients who eventually experienced AL, had developed the condition. Furthermore, the BoW databases \mathbf{X}_{+4} and \mathbf{X}_{+4}^{bin} were created, where “+4” indicates that this BoW is based on all journal notes up to and including post-operative day four.

Table 4.4 shows CM , Se , Sp , and the number of selected features for all the considered databases. The area under the curve was also explored, but similar results were obtained.

Note that discriminatory power is revealed, even for \mathbf{X}_{op} and \mathbf{X}_{op}^{bin} . In particular, for \mathbf{X}_{op}^{bin} , the results show that given that the patient will eventually experience AL, our FS method detects that in 100% of the cases. On the other hand, given that the patient does not eventually experience AL, our FS method correctly reveals that in about 72% of the cases. This means that the FS approach advocated in this application, has capacity for detecting AL patients at an early stage. Note also that the number of features selected in order to achieve these results is dramatically lower than the cardinality of the input feature space. As one would expect, the discriminatory power in the data increases with time.

Table 4.5: Words associated with selected (bootstrap) SVM positive weights corresponding to \mathbf{X}_{op}^{bin} (first column) and \mathbf{X}^{bin} (second column).

\mathbf{X}_{op}^{bin}	\mathbf{X}^{bin}
anastomosis	anastomosis leakage
shaved	anastomosis
easy	re-operated
relieving	re-operation
low	butt
localized	insufficiency
air	saline
info	anterior
up	vatan
anterior	colorectal
peripheral	some
far	drainage
anesthesia	sigmoidostomi
evt	suture
stapler	stapler
loop ileostomy	furix
coloanal	localized

Interpretation of Selected Words

One of the major advantages of training a linear SVM on the EHR is that each weight in the weight vector \mathbf{w} corresponds to a particular word in the BoW database, enabling knowledge extraction by analyzing the weights. In this section, those words that correspond to the dominant SVM weights are presented, and interpreted the words in the context of AL detection.

The databases \mathbf{X}_{op}^{bin} and \mathbf{X}^{bin} are analyzed in detail due to the promising AL detection results presented in the previous section. These databases contained only positive elements (binary numbers), such that a positive weight corresponded directly with the positive class (AL) and a negative weight was associated with the negative class, since the classification into the positive or negative class is based on the sign of $\mathbf{w}^\top \mathbf{x}$.

Those weights with the largest positive values correlate the most with the positive class, and vice versa for the negative class. Table 4.5 (first column) shows the words corresponding to some of the largest positive weights (in order) for \mathbf{X}_{op}^{bin} . These were the words which SVM associates with the positive class, i.e., the class of patients experiencing post-operative AL. For some surgeons in the University Hospital of North Norway, the appearance of several of these words in association with AL seemed reasonable from a clinical perspective. Some examples are presented below.

Tumors located in the lower part of the rectum are known to increase the risk of AL, and

4.2 Application 4.1: Early Detection of Anastomosis Leakage from Bag-of-Words93

are removed by the surgical procedure known as *low anterior resection*. Similar reasons may explain the appearance of the word *anterior* in Table IV (first column). The word *air* may be an indicator of a leakage, since the presence of air outside of the bowel will be due to a leakage. A diverting *loop ileostomy* will be performed in patients with the highest risk of AL (low rectal cancer with *coloanal* anastomosis and after neoadjuvant treatment with irradiation).

Some examples regarding the words associated by the SVM to the negative class, are the following: *amputation*, *abdominoperineal*, *endcolostomy*, *proximal*, *sonor percussion sound*. One of the words is *amputation*. This word simply refers to the removal of the whole rectum and anus in order to remove a distal rectal tumor oncologically safe. In that case the problem of AL is completely avoided and the patient will have a permanent endcolostomy. Abdominoperineal amputation is the name of the operation. Patients with *proximal* (means located in the upper part of rectum) rectal cancer do not need deep pelvic surgery and are thereby less exposed to AL. The expression *sonor percussion sound* is used when the physician describes the normal sounds that appear before the operation, when he/she carefully knocks on a finger placed on the patient's chest in order to detect pleural fluid collections or abnormal air distribution in the chest. AL is often followed by lung and heart complications.

We also analyzed the selected (bootstrap) SVM weights corresponding to the databases \mathbf{X}_{+4}^{bin} and \mathbf{X}^{bin} . The distribution of positive and negative weights change for these databases, compared to \mathbf{X}_{op}^{bin} . We focus here on the words corresponding to \mathbf{X}^{bin} . Table 4.5 (second column) shows the words corresponding to some of the largest positive weights (in order) for \mathbf{X}^{bin} . Several of the words from Table 4.5 in the first column reappear in the second column. However, there are differences. For example, the weight associated with the word *anastomosis leakage* is now the largest of all the weights. Furthermore, words like *re-operation* and *re-operated* are also associated with large weights.

This analysis shows that the selected words, obtained by the proposed FS strategies based on the BoW model, may be reasonably interpreted in the medical context of AL. Future work may consider highlighting words of particular medical relevance when training decision support systems, or flag certain selected words as indicators of the AL complication.

4.2.4 Discussion and Conclusions

In this work, it is demonstrated that the clinical narrative contains relevant information for early detection of AL following surgery for CRC. The discriminatory power in the data is based on a simple BoW model, where classification and FS is based on a linear ν -SVM.

Results show that both binary and non-binary approaches have discriminatory power. A binary input space yields a sensitivity of 100% and specificity of 72% at an early stage, while performance worsens when using a non-binary input space, to 97% and 47% respectively.

The number of relevant features is also lower in the latter case. In multidisciplinary studies like the present one, validation by clinicians is highly necessary in order to extract correct and useful knowledge. The set of words shown in Table 4.5 was therefore validated by a group of surgeons who concluded that several words (in bold) appear to have relevance for identification of patients with increased risk of AL after CRC surgery.

The study has some limitations. In particular, the number of cases is low, and hence prone to over-fitting, such that external validation of the results would be desirable. A manual annotation process as used here is likely to provide accurate labels, but is very time consuming. By using automated phenotyping [142] of the EHR, one can effectively gather larger cohorts at the loss of some accuracy. The extracted text does not contain all information about the patient, and notably the surgery notification form is unavailable. Thus there is much information missing about patient's preoperative status, which could be important additional indicators of subsequent complications and could improve accuracy.

In studies of risk assessment models there is always the concern that the signal may be censored when a clinician spots a pattern leading to a complication and takes appropriate action to avoid the complication [143]. This would result in a significant number of cases where the pattern leading to the adverse event is present but not the event itself as that was successfully averted, effectively constituting mislabeled cases. This might be a concern in the current study, and would, if the classifier is good, lead to a decreased specificity. In the work a BoW model was used, which is arguably the simplest possible model for text processing. Nevertheless it was demonstrated potential for FS for improving the AL detection.

This innovative study describes the development of an early computerized warning system that, when fully developed, will be a useful supplement for the clinician to be alerted at an early stage and act promptly to avoid potentially lethal post-operative outcomes. It is important that the information provided by the system is actionable on the part of the physician, in that there is an option to change the course of action for patients with increased risk. In the case where the risk is evident prior to the index surgery, potential courses of action can be to postpone the surgery until all known risk factors are corrected or to protect the anastomosis by a diverting stoma or avoid any anastomosis by giving the patient a permanent stoma in the first place. Additionally, the patient can be involved in the preoperative decision-making and sign an informed consent form based on a better understanding of the preoperative risk-assessment for AL. In the case of increased risk post-operatively, potential actions in the case of alarm signals indicating an anastomotic leakage would be emergency CT scans, diagnostic laparoscopy, or laparotomy. The latter two are resource demanding and not without potential complications. It is therefore beneficial to have additional computerized algorithms as described in this application, in addition to sound clinical judgment.

We have shown that there is information in the clinical narrative that can be used to predict AL after CRC surgery. Thus, the text can be a piece of the input to a clinical information system that warns clinicians of the potential for complications in individual patients. Experimental results corroborate the feasibility and sustainability of the proposed framework, although future work could further enhance results to support early diagnosis decisions.

4.3 Application 4.2: Early Detection of Anastomosis Leakage using Heterogeneous Sources

4.3.1 Introduction

In the previous application, one of the data sources, namely the free text clinical narratives, discovered a potential for detecting AL after CRC surgery based on this source [144, 145]. However, several works in different contexts have concluded that the combination of heterogeneous sources of information enhances classification and regression results in many applications, such as intelligent transportation systems [146], multibiometric face recognition application [147], and remote sensing [148]. The combination of heterogeneous sources from EHRs have been only moderately studied in the literature, likely due to the fact that the availability of the EHR information is limited, in some cases for privacy issues. However, some previous works concluded that the use of heterogeneous data improved clinical decisions. For example, the combination of structured EHR data (diagnostic codes, vital signs etc.) combined with free-text analysis in order to detect acute respiratory infections was analyzed in [29], enhancing sensitivity values in unhealthy patients. Merging heterogeneous clinical data from five databases improved Alzheimer's diagnosis [28].

In the ML literature, the so-called *composite kernel methods* have been used for combining heterogeneous sources in several applications [59]. For example, the task of hypertext categorization exploring words and links information individually and by using composite (combined) kernels was analyzed in [149], obtaining better performance by a combined kernel approach. Composite kernels were also used for hyperspectral image classification [148] and for the classification of very high resolution urban images [150]. Regarding clinical applications, the use of the composite kernel framework provided the highest classification rate for diagnosis of cancer based on colon cancer and leukemia datasets in combination with proteome patterns of a stomach cancer dataset [27], for the improvement of Alzheimer's diagnosis [28], and for the automatic diagnosis of pathological myopia [33], among others.

In this application, a prediction model based on structured and unstructured clinical data from the EHR is proposed for early detection of AL. The novelty of the present work is found in: (1) the exploitation of heterogeneous data sources for AL detection; (2) the leverage of the

power of composite kernels for this purpose; and (3), the assessment of a temporal risk score in order to control the patient status and detect AL complication at an early stage.

4.3.2 Database

In this section, three different data sources (free text, blood tests, and vital signs) are presented, which have been jointly analyzed in order to perform early detection of AL. First, it is explained how these data sources were recorded in the EHR and the specific characteristics of each data set. Later, the extraction and preprocessing stages needed to obtain a suitable input space to be treated by the classifiers proposed in this work are discussed. The specific nature of each data source required the development of different preprocessing strategies.

The same database described in the previous application, Application 4.1, is also used in this one, although more clinical documents were considered. In the current study, data from two heterogenous sources, namely, blood tests and vital signs, are also analyzed. A summary of the new sources is given next.

Blood tests. In this work, structured data from nine different laboratory tests, namely, albumin, C-reactive protein (CRP), glucose, hemoglobin, potassium, creatinine, leukocytes, sodium, and thrombocytes were analyzed. These blood tests were recorded for a period of 10 days before the surgery and up to 20 days after the surgery.

Blood tests measurements are in general highly irregularly extracted in time. This is illustrated in Figure 4.3 for the CRP blood test, showing that available data are characterized by a strongly irregular time sampling pattern. Hence, the observed data matrix is *sparse* over patients and time. This poses challenges in the data processing. From a data processing perspective, the data sparseness is equivalent to missing data, and must be handled. The irregular sampling and resulting sparseness of data is even higher for some of the other tests (not shown here).

Despite the efforts made to develop statistical methods for handling missing data, there is no global best approach because the different approaches depend on different assumptions. When a relatively small number of samples are missing, skipping features or patients can be an option, but this was not the case in our problem. An imputation method based on the k -NN algorithm as in [43] was followed, which allowed us to work with a database denoted as \mathbf{X}^{blood} from now on.

Vital signs. Three vital signs (temperature, blood pressure -high and low values-, and pulse) were extracted from different types of nurse's notes using several layered regular expressions working on a specific part of the different documents where this information was noted down. Vital signs were normally recorded at least three times per day for each patient, for a period of 10 days before the CRC surgery up to 20 days after the surgery. Since these data are by nature irregularly sampled, thus, an imputation method based on the k -NN algorithm was applied to

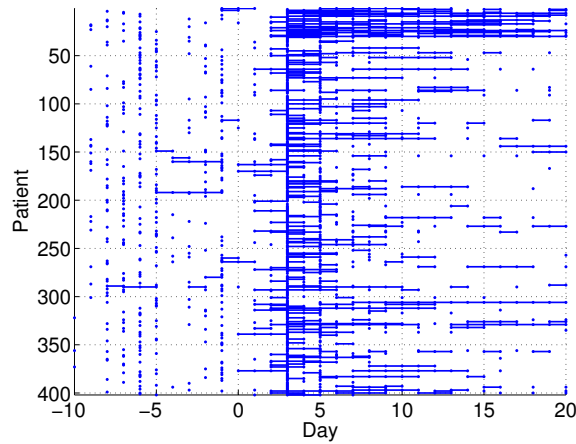


Figure 4.3: *CRP laboratory test measurements for each patient (the first 31, from the top, correspond to AL cases) for a period of 10 days before the colorectal surgery and up to 20 days after the surgery (Day 0). The data matrix is sparse over patients and time, constituting a challenging data set to work on.*

obtain daily measures.

Temperature. The extraction process restricts it to be between 30.0 and 41.0 as normal values. In some cases, the analyzed text contained words indicating that the patient was afebrile, febrile or subfebrile.

Blood pressure. It measures the diastolic and systolic blood pressure of a patient, given as two integers separated by a /, that is for instance 120/80. The extraction process restricts it to be: (1) first integer larger than second integer; (2) first integer (overpressure) larger than 60 and lower than 250; and (3) second integer (underpressure) larger than 30 and lower than 200. The analyzed free text sometimes contained words telling that the patient had normal, low, or high blood pressure.

Pulse. The number of heart beats per minute was given as an integer. The extraction process restricts it to be between 41 and 250. Choosing 41 as the lower limit makes sense medically, though there might be rare cases of lower pulses than this. In these cases, the patient was probably anyway kept under tight control. Free text sometimes contained words indicating that the patient had normal pulse, irregular heart beat, or irregular heart beat with an approximate number of beats per minute given in the text. Only a really small number of patients had irregular beats, so this characteristic was discarded and only was analyzed the number of heart beats per minute.

Data from vital signs were represented as a concatenation of four values (temperature, high and low pressure, and pulse), in matrix \mathbf{X}^{signs} from now on.

4.3.3 Experiments and Results

In this section the individual contribution of each data source is presented, as well as the results obtained after combining heterogeneous sources using composite kernel. The experimental setup followed in this application is the same as the one described in the previous one. Thus, it is not repeated here.

Individual Contribution of Each Data Source

The potential to predict AL based on data from individual sources recorded in the EHR was explored. Towards this end, the performance of linear and non-linear SVM classifiers was evaluated. First, we automatically selected the free parameters following a LOO strategy. The linear ν -SVM method requires the tuning of a single free parameter, $\nu \in (0, 1)$. For the RBF-based ν -SVM, the tuning of the width σ is also necessary. Furthermore, results after evaluating RFE and bootstrap FS methods were provided. The first one was computed for linear and non-linear SVM classifier, whereas the second was only computed for the linear case.

Free text. A BoW model analysis based only on data up to and including the day of the surgery identified AL patients with $Se = 97\%$ and $Sp = 66\%$ using a linear kernel after a FS strategy (see Table 4.6). Note that the performances enhanced when more information was considered ($Se = 100\%$, $Sp = 68\%$ for \mathbf{X}_{+4} , and $Se = 100\%$, $Sp = 87\%$ for \mathbf{X}). A linear SVM classifier considering the features subset after a bootstrap resample strategy provided the best predictions when only free text data set was evaluated. A RBF SVM classified all patients in the same class. For this reason, and in order to avoid the high dimensional input space in the BoW model, we decided to focus only on the features subset obtained after considering a bootstrap FS method from now on.

Blood tests. For this data source, Se and Sp improved when a non-linear SVM classifier was considered. More specifically, the application of the non-linear RFE FS strategy enhanced the performance for all time slots.

Physiological data. We also evaluated to what extent AL can be detected based only on vital signs. A linear classifier did not performance properly, but non-linear SVM classifiers provided reasonable Se and Sp values. Results improved when using a non-linear RFE FS method.

In summary, the linear SVM performed the best when a BoW model was considered, yielding $Se = 97\%$ at an early stage. However, a higher Sp was obtained when blood tests data for \mathbf{X}_{op} were analyzed. As indicated in Application 4.1, the clinical narrative provided the best performance when all available data were considered separately.

Table 4.6: Classification sensitivity (first value, in %), specificity (second value, in %), and number of features (in brackets) for linear and non-linear kernels when individually applied on free text data set, blood tests and vital signs. Best values are shown in bold.

BoW	\mathbf{X}_{op}		\mathbf{X}_{+4}		\mathbf{X}	
Linear kernel	65	58 (5409)	74	68 (6858)	84	85 (13188)
Linear & boot	97	66 (0158)	100	66 (0186)	100	87 (00389)
Linear & RFE	65	58 (5406)	74	68 (6476)	84	85 (13180)
Blood tests	\mathbf{X}_{op}		\mathbf{X}_{+4}		\mathbf{X}	
Linear kernel	71	74 (99)	77	72 (135)	77	76 (279)
Linear & boot	68	76 (96)	71	70 (128)	77	77 (186)
Linear & RFE	71	74 (98)	81	73 (108)	77	75 (234)
RBF kernel	87	60 (99)	87	71 (135)	94	63 (279)
RBF & RFE	87	68 (93)	90	72 (130)	97	77 (040)
Vital Signs	\mathbf{X}_{op}		\mathbf{X}_{+4}		\mathbf{X}	
Linear kernel	61	20 (44)	61	39 (60)	52	49 (124)
Linear & boot	55	29 (42)	45	41 (55)	65	40 (118)
Linear & RFE	61	25 (33)	42	42 (52)	52	41 (108)
RBF kernel	68	65 (44)	61	56 (60)	94	52 (124)
RBF & RFE	65	62 (33)	68	48 (21)	81	71 (093)

Heterogenous Data Sources Combination

Previously, it was concluded that free text provides the higher Se values. In this section, it is analyzed whether clinical results can be improved by combining different data sources available in the EHR. Towards that end, the classification performance when using a stacked kernel and a composite kernel method were benchmarked. Results are shown in Table 4.7.

Stacked input vectors kernel. Features from pairs of two sources were first stacked in an input vector using a single kernel, obtaining three different combinations, namely, BoW with blood test, BoW with vital signs, and blood tests with vital signs. Then, all features from the three sources were stacked in an input vector, using a single kernel. In this case, the new vector dimension is obtained as the sum of the three datasets dimensions, evaluated for \mathbf{X}_{op} , \mathbf{X}_{+4} , and \mathbf{X} , respectively. After the new input vector was built, linear and non-linear classifiers were designed, both with and without a FS strategy.

Results showed improved AL detection in general when combining heterogenous sources. In general, clinical narrative had good discriminatory power itself, and also after combining it with blood tests or vital signs. More specifically, free text and vital signs fusion was promising in order to detect AL at an early stage ($Se = 100\%$ and $Sp = 72\%$). Using only structured data provided some reasonable classification results, however, they were inferior compared to those obtained when considering free text.

Table 4.7: Classification sensitivity (first value, in %), specificity (second value, in %), and number of features (in brackets) for linear and non-linear kernels when combining free text, blood tests and vital signs data sources. Best values are shown in bold.

BoW & Blood	X_{op}		X₊₄		X	
Stacked (linear)	97	68 (257)	100	69 (321)	100	87 (668)
Linear & boot	77	61 (244)	97	66 (300)	100	85 (615)
Linear & RFE	97	68 (256)	100	69 (318)	100	87 (668)
Stacked (RBF)	90	63 (257)	97	59 (321)	68	42 (668)
RBF & RFE	77	71 (093)	97	73 (226)	100	87 (614)
BoW & VS	X_{op}		X₊₄		X	
Stacked (linear)	93	63 (242)	100	70 (250)	100	88 (513)
Linear & boot	90	64 (227)	97	68 (243)	100	86 (482)
Linear & RFE	97	64 (239)	100	69 (243)	100	88 (443)
Stacked (RBF)	90	70 (242)	97	57 (250)	68	46 (513)
RBF & RFE	100	72 (238)	100	74 (249)	100	88 (457)
Blood & VS	X_{op}		X₊₄		X	
Stacked (linear)	58	73 (143)	61	76 (195)	77	73 (403)
Linear & boot	54	69 (140)	65	75 (189)	71	71 (333)
Linear & boot	64	57 (142)	58	59 (188)	81	74 (360)
Stacked (RBF)	81	74 (143)	65	83 (195)	81	73 (403)
RBF & RFE	87	80 (125)	84	68 (187)	87	77 (392)
BoW & VS & Blood	X_{op}		X₊₄		X	
Stacked (linear)	97	67 (301)	100	75 (381)	100	87 (792)
Linear & boot	84	67 (306)	100	74 (372)	100	86 (755)
Linear & RFE	97	66 (250)	100	73 (165)	100	88 (691)
Stacked (RBF)	90	57 (301)	87	56 (381)	68	41 (792)
RBF & RFE	100	69 (277)	97	77 (327)	100	87 (730)
Composite kernel	100	76 (301)	100	73 (381)	100	88 (792)

Composite kernels. It was evaluated whether results improved by exploiting a combination of different kernels for all the available sources. After analyzing individually the three data sources, the conclusion was that the best classifier scheme for free text data were a linear SVM, whereas for blood test and vital signs data sets, a non-linear scheme enhanced detection performances. Thus, the used composite kernel can be expressed as:

$$K = K^{blood} + \mu_1 K^{BoW} + \mu_2 K^{signs} \quad (4.1)$$

where K represents the composite kernel, K^{BoW} is the linear kernel based on BoW model, and K^{blood} and K^{signs} represent the RBF kernel when considering blood test and vital signs data sets, respectively. First, the kernel free parameters were tuned, and then, a LOO cross-validation optimization procedure was developed to select μ_1 and μ_2 values, obtaining the optimized

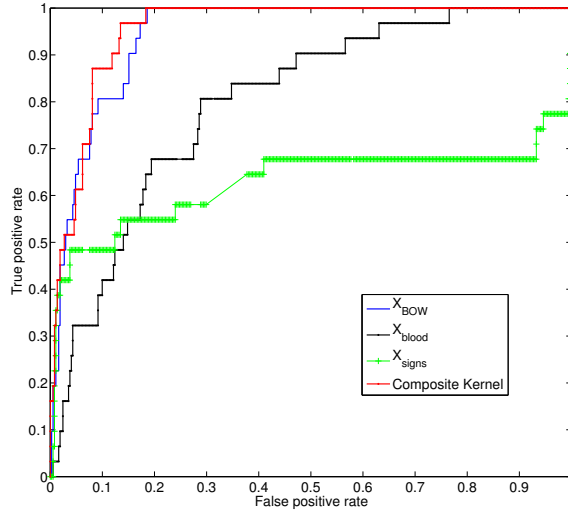


Figure 4.4: ROCs using X_{op} and individual data sources X^{BoW} , X^{blood} , and X^{signs} with ‘Linear & boot’, ‘RBF & RFE’ and ‘RBF kernel’ classifiers, respectively. ROCs using X^{op} and combination of the three individual data sources with the composite kernel in Eq. (4.1).

composite kernel. The search ranges for μ_1 and μ_2 was (0,10).

Composite kernel results are shown in the very last row of Table 4.7. Note how already at the day of surgery, sensitivity and specificity are at 100% and 76%, respectively, clearly indicating an improved capability for early detection of AL. This is promising, and it shows that composite kernels have a potential for extracting useful information from the heterogeneous data sources which are considered in this work.

For a visual interpretation, the Receiver Operating Characteristic (ROC) was represented. The ROC curve is generated by varying the threshold parameter in the classifier which controls the trade-off between sensitivity and specificity. In this case, the soft output for yielding a statistical decision parameter on which moving the threshold was used. Figure 4.4 shows individual ROCs for X^{BoW} , X^{blood} , X^{signs} and after combining them using a composite kernel when considering data up to and including the day of the surgery. For X^{BoW} , the ROC was calculated after considering a linear SVM classifier with a bootstrap FS method, whereas for X^{blood} and X^{signs} , a non-linear SVM and a RFE FS method were considered.

Temporal Risk Assessment for Early Detection of AL

Previous experiments showed the AL prediction performance for the 402 patients at different time slots. However, a more complete patient risk assessment may be useful. For example in order to warn clinicians to be alerted at an early stage and act promptly to avoid complications.

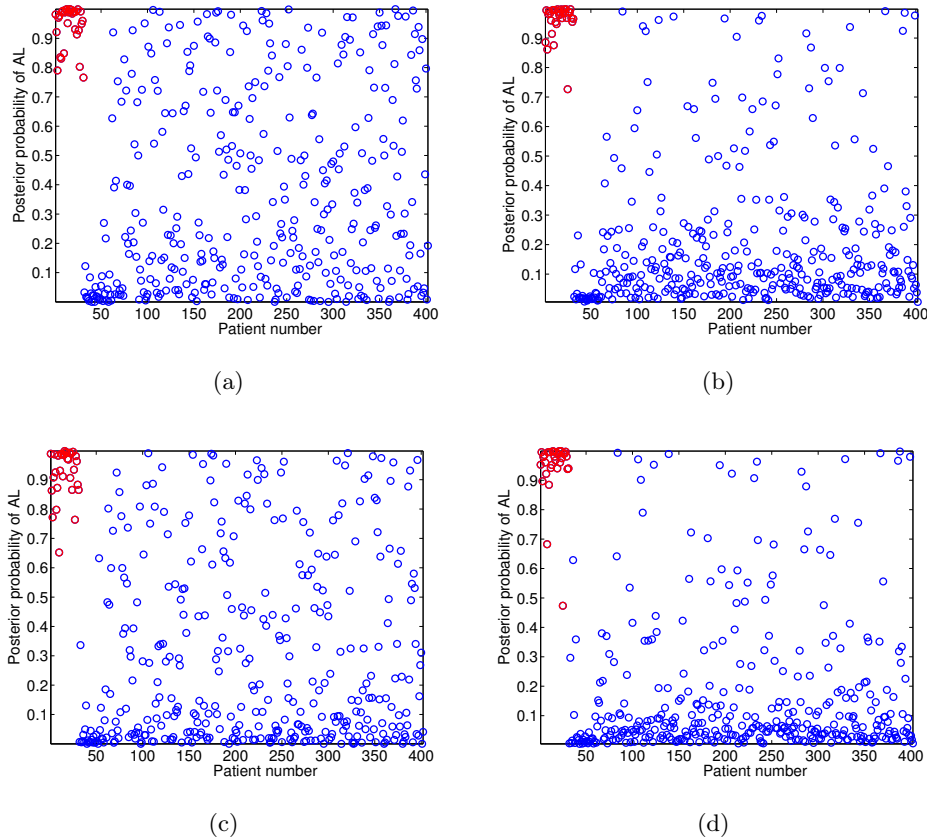


Figure 4.5: Estimation of posterior probabilities for each patient (infected in red, not infected in blue). Upper panels: using \mathbf{X}^{BoW} and linear SVM with FS for \mathbf{X}_{op} (a), and \mathbf{X} (b). Lower panels: using the three heterogeneous sources and a composite kernel for \mathbf{X}_{op} (c), and \mathbf{X} (d).

When the risk is evident, potential courses of action can be applied to avoid AL. Towards that end, we proposed a temporal risk score using several heterogeneous data sources. Following the same two-stage procedure, as explained in Sec. 4.2.3, the risk score is based on the estimated posterior probabilities of AL, obtained after training the SVM classifier, i.e, given the i -th patient with feature vector \mathbf{x}_i , the goal is to estimate $Pr(y = 1|\mathbf{x}_i)$, for $i = 1, \dots, n$ (see Sec. 2.3.1). The higher/lower (infected/not infected) the posterior probability is, the higher the confidence in the classification will be, and thus, improving the likelihood of better clinical decision support.

Figure 4.5 shows the posterior probability estimation for each patient; red points are associated to infected cases (positive class), while blue points correspond to negative cases. Fig. 4.5 (a) and Fig. 4.5 (b) represent the estimated posterior probability when only free text is considered to predict AL, both using \mathbf{X}_{op} and \mathbf{X} , respectively. In order to compare these results with those obtained when several heterogeneous sources are considered, the estimated posterior probability for \mathbf{X}_{op} and \mathbf{X} (Fig. 4.5 (c) and Fig. 4.5 (d)) is represented. Note that, in general,

lower posterior probabilities of AL are reported for the negative cases (controls) when using more sources.

4.3.4 Discussion and Conclusions

In this work, we demonstrated that the combination of heterogenous data sources (i.e. free text, blood tests, blood pressure, pulse and body temperature) from the EHR using SVM and composite kernels may predict AL at an early stage. Free text, laboratory tests, and vital signs, have been simultaneously used in order to deal with both unstructured and structured data. The impact of each data source individually, as well as their combination using weighted kernel summations, were studied.

Our results show that improved classification performance was obtained when clinical narrative was used. Nevertheless, the individual analysis of laboratory tests and vital signs also provided with additional discriminatory power. An increase in sensitivity classification is clearly obtained when combining these heterogenous data sources. Furthermore, the risk assessment status of the patient improved when using multisource information from the EHR. This result is specially relevant to detect AL complications at an early stage, and its inclusion in an on-line prediction system might be used to predict patients risk of AL.

For the sake of simplicity, an imputation method based on the k -NN algorithm was used to deal with missing data, though other imputation strategies will likely yield more accurate classifiers. Different methods to deal with observations at non-uniform time points have been grouped in three different categories [151]: (1) smoothing or interpolation techniques to fill missing observations; (2) spectral analysis tools, such as wavelets or Lomb-Scargle Periodogram; and (3) kernel methods. All of them depend on the considered assumptions on the data, and are sensitive to the time series dynamics. Furthermore, for the cases with several heterogenous data sources recorded with different criteria, more specific and elaborated methods have to be developed to this end. On the other hand, we only used a state-of-art non-linear FS method in this work, but more theoretical and experimental work should be devoted to the topic of large input spaces in the EHR data sources. Note also that sample imputation and FS are strongly coupled problems in this kind of sparse temporal data, so it is recommendable to develop methods for their joint assessment. Finally, we used state-of-art posterior probability estimation, and given the suitability of this type of output in the clinical environment and the improvement of the detection capabilities in our results, further theoretical and experimental effort is also encouraged in this setting.

The suitability of vital signs to diagnose an AL after intestinal resection was analyzed in [152], showing that it represents a quite challenging problem, and low prediction capabilities were obtained therein. However, CRP showed promising diagnostic value in excluding patients without

AL [153]. Furthermore, both vital signs and laboratory tests data were also combined into a clinical score system [154], in order to identify patients with higher risk of AL using statistical tests, such as Mann-Whitney and ANOVA, hence improving the prediction. In this work, and based on kernel methods, it was able to combine vital signs and laboratory tests data together with free text, yielding higher discriminative power.

4.4 Application 4.3: Data-driven Temporal Prediction of Surgical Site Infection

4.4.1 Introduction

When using observational data from secondary sources such as the EHR one needs to take into account that the information is rarely recorded in a systematic way. Indeed, the data are often sparse, and gathered at a clinician's discretion. For example, blood tests are taken at a mixture of predefined stages in a patient pathway and clinically driven sampling. Thus, if predictive analytics relies on regularly sampled data, imputation methods need to be employed such that regular sampling is simulated. However, in the case of very irregular sampling a classical imputation approach may not be sufficient. In this work, prediction models are studied for real time evaluation of patients admitted for gastrointestinal surgery with respect to surgical site infection (SSI) post-operatively.

The American College of Surgeons Surgical Quality Improvement Program and the Centers for Disease Control and Prevention divide SSI into three subtypes based on the anatomical location of the infection, i.e. superficial, deep incisional and organ space [126]. Superficial infections can usually be cured with per oral antibiotics and surgical debridement. In contrast, deep and organ space SSI require intravenous antibiotics, percutaneous drainage and laprotomies.

The patient specific risk factors for SSI are well documented and reported. A recent study by Lawson et al [126] identified open surgery, ulcerative colitis, older age, overweight, smoking, disseminated cancer and prolonged operation time as factors contributing to an increased risk of SSI. However, they found that different risk factors were associated with superficial and deep SSI. High body mass index and revision of an osteomy were associated with superficial SSI, whereas prolonged operation time and perioperative transfusions were associated with organ space SSI [126, 128].

Using blood test results as predictive features in a data-driven decision support system is useful since these are performed relatively often with little burden to the patient. Therefore it is, e.g., possible to estimate the expected information content of a blood test at stages in a patient trajectory [155]. However combining different tests performed at different stages in the trajectory, which is necessary when observational data are used, presents challenges which are

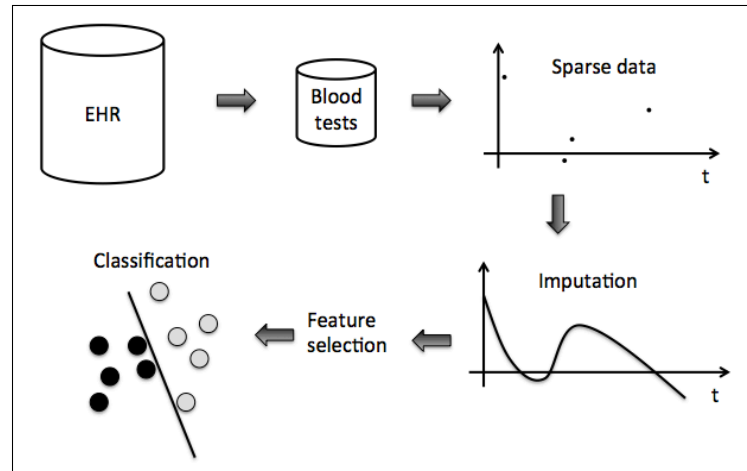


Figure 4.6: Overview of the processing pipeline.

Table 4.8: Demographic characteristics of the patient groups.

	Overall	Controls	Cases
Female (%)	477 (47.4)	441 (48.7)	36 (35.6)
Age [Mean \pm SD]	57.0 \pm 20.7	56.9 \pm 21.2	57.4 \pm 15.2

addressed here. The information from tests may be further combined with other data such as textual features that are predictive of complications [145].

For the purpose of this Thesis, we denote the sparsity of the clinical data as missing data. Missing data percentages are even larger for some studies such as clinical laboratory measurements or biomarkers. Despite of the efforts made to develop statistical methods for handling missing data, there is no global best approach because they inevitably depend on stated assumptions.

In this work methods for predictive modeling in a context of features that have strongly irregular sampling patterns are presented. Different smoothing and interpolation/imputation techniques and different input spaces to predict SSI using blood tests are analyzed. Finally linear and non-linear classifiers are computed to do the predictive modeling. Figure 4.6 shows an overview of the data-driven decision support system used in this work for SSI prediction.

4.4.2 Database

A cohort of patients based on relevant International Classification of Diseases (ICD10) or NOMESCO Classification of Surgical Procedures (NCSP) codes related to severe post-operative complications, and in particular to SSI, was extracted from the EHR of the department of Gastrointestinal surgery at the University Hospital of North Norway. The selection of codes was guided by input from clinicians at the hospital. The cohort identified as control was

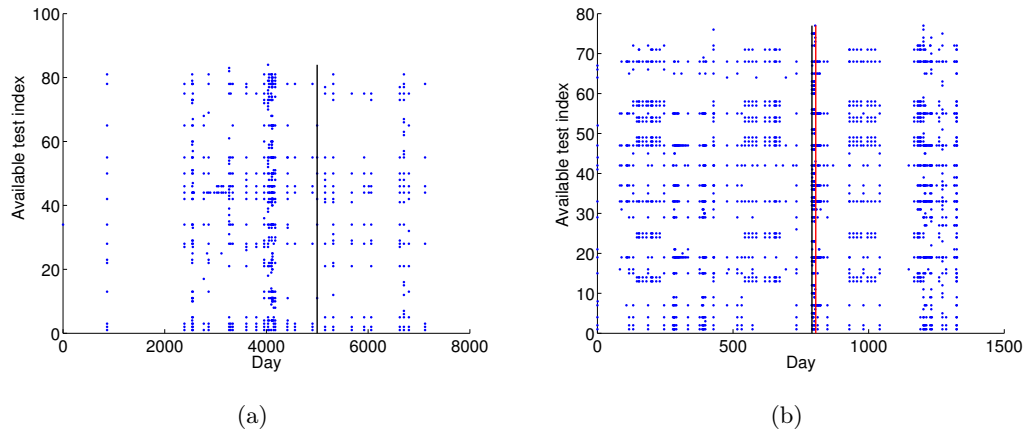


Figure 4.7: Available laboratory test measurements for one control (a) and another infected (b) patient. The y-axis shows the available tests for a patient, with no specific order, as each patient can have different number of tests. The x-axis represents the day when each test was recorded, being Day = 0 when the first test was recorded. Vertical black line indicates the surgery day, whereas red line indicates the infection day.

matched with patients that did not have any of these codes but were otherwise similar in terms of which blood tests were performed. Additionally, a text search was performed to ensure that the controls did not have the word “infection” in any of their post-operative text documents. This resulted in a cohort of 101 cases and 904 matched controls. Patients with codes indicating superficial infections were excluded. A set of 10 different types of blood tests, namely, hemoglobin, leucocytes, CRP, potassium, sodium, creatinine, alanine aminotransferase (ALAT), thrombocytes, albumin and alkaline phosphatase (ALP), were defined as clinically relevant and extracted for all patients from their EHRs. All tests were not available every day, which results in a high percentage of missing values when analyzing data on that scale, yielding to a non-uniform time sampling description for each patient (Fig. 4.7). The data matrix is hence sparse over lab tests and time, therefore constituting a challenging data set to work on. The proposed method in this Dissertation, denoted as bootstrap nonparametric resampling, was designed to statistically describe the influence of imputation (see Sec. 2.5.2). Thus, the population mean and corresponding 95% CI was computed on a daily basis for each test, obtaining an averaged trend.

The data represent a diverse group of patients undergoing gastrointestinal surgery such that results can generalize across this group. The basic demographics of the cohort are given in Table 4.8.

4.4.3 Experiments and Results

Predictive Analytics of SSI

Feature engineering for sparse clinical data. Working with complete datasets is the standard scenario for most statistical and ML methods. In the literature, there are works that simply omit patients with any missing data, but it is not a reasonable approach with high-dimensional data. To avoid this situation, different methods have been proposed to deal with observations at non regular sampling. These methods can be categorized into: [151] (1) smoothing or interpolation techniques; (2) spectral analysis tools such as wavelets or Lomb-Sargle Periodogram; and (3) kernel methods.

Regarding interpolation methods, the well-known Last Observation Carried Forward (LOCF) scheme imputes the last non-missing value for the following missing values [156]. Alternatively, Lasko et al. [44] suggest using GP followed by a warped function [44], and this approach is followed in this work. The warped function explained in Sec. 2.2.1 is intended to adjust for the fact that rapid changes in temporal variables in connection with active treatment is often followed by long periods of apparent stability leading to highly nonstationary processes. This function converts non-stationary clinical data into a stationary process which allows the use of a GP (see Sec. 2.3.1 for details) to deal with sparsity. Thus, we use this approach in this manuscript to deal with sparse data.

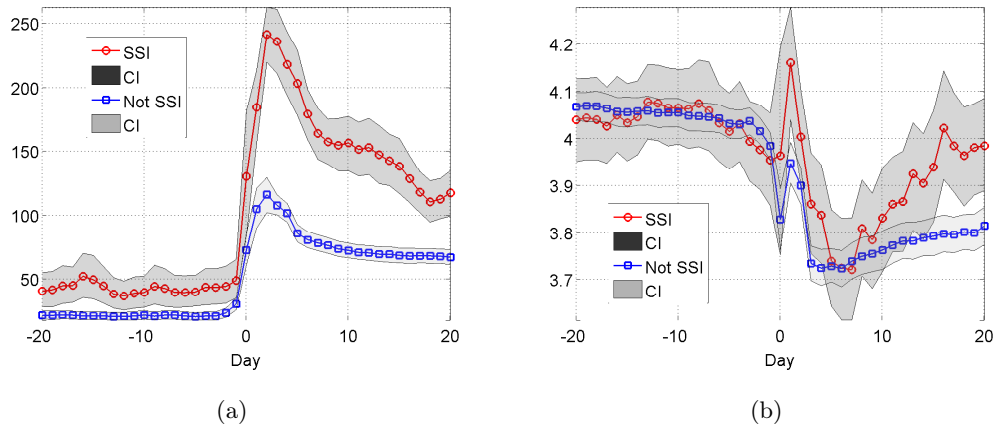


Figure 4.8: Population mean and corresponding 95% CI per day for CRP (a) and Potassium (b) tests, being Day = 0 the day of the surgery. LOCF imputation and a nonparametric resampling strategy have been used.

Experimental Setup. First, the capabilities of different ways to deal with sparse data and to show the effects on performance results are evaluated. Furthermore, linear and non-linear classifiers are benchmarked to predict SSI when using different laboratory tests data obtained

Table 4.9: Pre-operative and post-operative accuracy (mean and 95% CI) for each test individually and different classifiers: Logistic regression (first row), linear SVM (second row), and non-linear SVM (third row). The best accuracy values for pre-operative and post-operative are shown in bold.

Lab test	LOCF		Warped-GP	
	Pre-operative	Post-operative	Pre-operative	Post-operative
Hemoglobin	0.48 [0.43,0.53]	0.47 [0.44,0.75]	0.60 [0.54,0.64]	0.60 [0.54,0.64]
	0.58 [0.50,0.69]	0.62 [0.51,0.69]	0.52 [0.40,0.62]	0.55 [0.46,0.63]
	0.70 [0.56,0.84]	0.89 [0.77,0.95]	0.71 [0.64,0.81]	0.79 [0.65,0.85]
Leucocytes	0.50 [0.43,0.56]	0.47 [0.43,0.51]	0.54 [0.48,0.59]	0.54 [0.48,0.59]
	0.50 [0.38,0.59]	0.61 [0.50,0.71]	0.45 [0.30,0.55]	0.53 [0.44,0.65]
	0.75 [0.62,0.85]	0.77 [0.65,0.85]	0.75 [0.61,0.87]	0.81 [0.73,0.93]
CRP	0.49 [0.44, 0.55]	0.48 [0.44,0.54]	0.62 [0.51,0.73]	0.44 [0.41,0.50]
	0.51 [0.43,0.60]	0.79 [0.71,0.87]	0.50 [0.39,0.67]	0.60 [0.47,0.71]
	0.61 [0.52,0.69]	0.90 [0.84,0.94]	0.79 [0.66,0.94]	0.79 [0.67,0.88]
Potassium	0.48 [0.44, 0.54]	0.47 [0.44,0.54]	0.52 [0.49,0.60]	0.48 [0.51,0.44]
	0.58 [0.49,0.66]	0.64 [0.46,0.72]	0.59 [0.52,0.69]	0.53 [0.63,0.43]
	0.73 [0.60,0.84]	0.88 [0.77,0.95]	0.66 [0.60,0.83]	0.74 [0.64,0.86]
Sodium	0.48 [0.44, 0.54]	0.47 [0.44,0.54]	0.49 [0.45,0.57]	0.48 [0.42,0.53]
	0.53 [0.43,0.68]	0.55 [0.34,0.73]	0.54 [0.42,0.70]	0.52 [0.46,0.58]
	0.66 [0.56,0.74]	0.76 [0.67,0.89]	0.71 [0.55,0.90]	0.68 [0.63,0.79]
Creatinine	0.46 [0.40,0.53]	0.46 [0.44,0.50]	0.49 [0.47,0.57]	0.41 [0.34, 0.45]
	0.55 [0.46,0.62]	0.61 [0.44,0.67]	0.50 [0.36,0.59]	0.52 [0.38,0.64]
	0.79 [0.73,0.86]	0.69 [0.56,0.82]	0.68 [0.55,0.74]	0.75 [0.69,0.83]
ALAT	0.50 [0.44, 0.53]	0.49 [0.44,0.53]	0.57 [0.49,0.64]	0.54 [0.48,0.58]
	0.61 [0.53,0.69]	0.54 [0.43,0.66]	0.63 [0.56,0.59]	0.49 [0.40,0.59]
	0.69 [0.50,0.82]	0.61 [0.47,0.71]	0.76 [0.63,0.88]	0.67 [0.63,0.75]
Thrombocytes	0.57 [0.48,0.63]	0.56 [0.47,0.62]	0.57 [0.49,0.65]	0.57 [0.54,0.60]
	0.56 [0.45,0.70]	0.66 [0.59,0.73]	0.61 [0.40,0.75]	0.49 [0.43,0.56]
	0.73 [0.62,0.83]	0.73 [0.66,0.89]	0.65 [0.58,0.70]	0.68 [0.58,0.74]
Albumin	0.53 [0.41,0.65]	0.50 [0.41,0.64]	0.56 [0.52,0.60]	0.47 [0.42,0.50]
	0.55 [0.40,0.66]	0.70 [0.44,0.84]	0.79 [0.55,0.92]	0.63 [0.54,0.69]
	0.71 [0.48,0.89]	0.82 [0.69,0.93]	0.91 [0.88,0.92]	0.83 [0.77,0.92]
ALP	0.49 [0.38,0.54]	0.49 [0.41,0.53]	0.41 [0.36,0.54]	0.33 [0.31,0.36]
	0.55 [0.43,0.67]	0.58 [0.53,0.65]	0.69 [0.64,0.75]	0.55 [0.44,0.71]
	0.69 [0.50, 0.84]	0.63 [0.47,0.76]	0.69 [0.44,0.87]	0.74 [0.69,0.79]

from the EHR. Firstly, each laboratory test was used separately to predict SSI using linear and non-linear classifiers after dealing with sparse data. Secondly, the use of multiple blood tests to check the impact of combining them as well as the temporal-feature relative importance was analyzed.

The database in this application was imbalanced, with 101 and 904 cases in the positive and

4.4 Application 4.3: Data-driven Temporal Prediction of Surgical Site Infection 109

negative classes, respectively. This is a common situation for clinical databases, where different number of patients are assigned to each class. Since previous studies have demonstrated that balanced classes in the training set often improve the overall classification performance [140]. An undersampling strategy (discarding samples from the majority class) was considered, such that the training set was built by enforcing balanced classes. In order to represent correctly the population, we selected a different number S of subsets of the negative class with 101 samples in each, and computed classification performances in terms of the mean and the standard deviation of the results for each subset.

A cross-validation strategy was used to ensure the generalizability of the prediction analytics. After balancing the classes, data were split into training and test subsets (80%-20%). A LOO cross-validation was carried out on the training subset of the balanced set for selecting the classifier free parameters.

Effect of the imputation methods on the performance. Two different strategies, namely, LOCF and warped-GP, were considered to deal with the extreme sparsity present in the input space as given by different tests measured in a patient at different days.

LOCF. The last observed non-missing value was used to fill in the missing values into a regular time sampling grid with a daily time basis, i.e., if there is a missing value, the previous value is considered if it exists. A nonparametric resampling method to represent the averaged trend was applied to statistically describe the influence of imputation. See two examples in Fig. 4.8 (a) for CRP and Fig. 4.8 (b) Potassium tests. It is well known that CRP is a good predictor for complications after colorectal surgery, and the pattern of CRP levels following surgery (see Fig. 4.8 (a)) is consistent with that observed by Singh et al. [157]. Note that the higher mean CRP levels before surgery reflects the smaller group size (cases) and thus larger variance in this case. For most blood tests, a wider CI after LOCF imputation was obtained for patients with SSI. Specifically, the data recorded at the day of surgery are highly noisy, as it can be seen in Fig. 4.8. For this reason, we excluded these values from our analysis, and we focused only on pre-operative and post-operative periods.

Warped function and GP. Using the time warped function Eq. (2.1), for each test we selected values of α and β parameters which maximize the accuracy of the predictive system. For this purpose, a grid search over values $\alpha \in [1, 10]$ and $\beta \in [0, 100]$ was evaluated. A LOO strategy was considered to ensure generalizability. The use of GP regression allows us to transform a set of finite measurements contained in the EHR from each blood tests into a continuous longitudinal function. In this way, missing values are inferred, allowing pre-operative and post-operative feature extraction.

Prediction of SSI. Table 4.9 shows the pre-operative and post-operative classification performance in terms of accuracy (mean and 95% CI) for each blood test individually when

considering a LOCF strategy and warped function with GP methodology. Pre-operative stage was defined as four days before surgery (i.e., $N = 4$), and four days immediately after surgery were considered in the post-operative stage (i.e., $N = 4$). Linear and non-linear SVM classifiers were considered for the prediction of SSI, and results were benchmarked with a simpler logistic regression classifier [158]. Results suggest the presence of strong non-linear relationship among input features for the analyzed tests, as given by consistently achieving the best performances when a non-linear SVM was considered. Note also that the post-operative predictive power is in general higher than pre-operative, which is to be expected.

Table 4.9 also shows that performance depends on the method used to deal with sparsity. In general, the combination of warped function and GP improved results, however, it can be seen that for some tests LOCF is better.

Feature Selection

Taking into account that nonlinear classifier provides a better prediction of SSI, the accuracy using a non-linear SVM classifier was obtained for both pre-operative and post-operative stages. First, all blood tests were considered together, i.e., $N = 40$ (first row in Table 4.10) and then the FS method denoted as RBF RFE was applied (second row in Table 4.10). Comparison of Table 4.9 and Table 4.10 shows that the model built with all tests provide in general higher accuracy. Note also that a similar or tending to higher accuracy is obtained with the FS method, so it is appropriate for addressing the interpretation of the relevance and meaning of the input space.

Figure 4.9 summarizes the results of FS with non-linear SVM (with RBF kernel) in terms of relevance of blood tests. Towards that end, how many times every feature is selected (frequency of relevance) was calculated, separately for the pre-operative and post-operative stages. From these values, a relevance index for each blood test is obtained as the normalization of the cumulative frequency of relevance by number of days ($N = 4$) times the number of subsets ($S = 5$). Note that a comparison with baseline level is remarkable for all tests (excepts sodium), indicating the relevance of the intra-patient pre-operative levels on each test. In general terms, thrombocytes reached the highest prediction information, together with ALP, CRP, albumin, creatinine and leucocytes, most of them being consistent with previous results. Although less relevant in the pre-operative state, the other tests (potassium, ALAT, and hemoglobin) also included highly relevant information in the post-operative state.

4.4.4 Discussion and Conclusions

The results clearly demonstrate the utility of blood tests for predicting SSIs both pre- and post-operatively. These results will potentially be useful as part of a data-driven online

4.4 Application 4.3: Data-driven Temporal Prediction of Surgical Site Infection 111

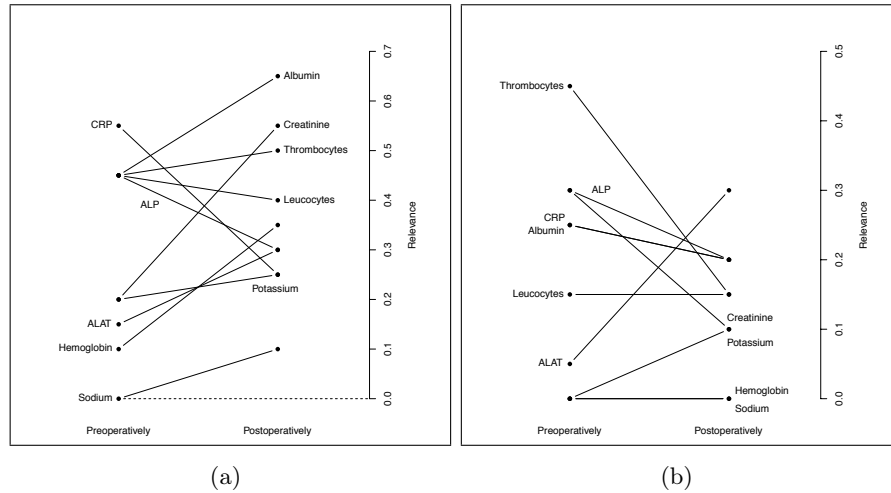


Figure 4.9: *Pre-operative and post-operative relevance index for each blood test using LOCF (a) and Warped-GP (b).*

Table 4.10: *Pre-operative and post-operative accuracy (mean and 95% CI), using all tests (RBF SVM) and after considering RBF RFE FS method.*

	LOCF		Warped-GP	
	Pre-operative	Post-operative	Pre-operative	Post-operative
All tests	0.81 [0.76,0.86]	0.89 [0.92,0.97]	0.88 [0.79,0.92]	0.90 [0.87,0.92]
FS	0.83 [0.67,0.90]	0.91 [0.90,0.92]	0.87 [0.76,0.94]	0.92 [0.90,0.94]

clinical decision support system that can enable clinicians to improve post-surgical recovery rates. With proper warning, necessary actions such as closer follow up and risk stratification can be performed.

Laboratory tests are often done at the discretion of the clinician, and often not driven by formulaic rules. This is part of the reason for the irregular sampling in the data, leading to the problem formulation in this application. Thus, the blood tests pattern for patients in inpatient care itself may be an informative feature of post-operative complications independent of the test results. By generating the cohort by matching blood tests patterns, this information is largely lost and only the test results remain as the informative features.

In retrospective EHR studies, there is inevitably the chance of a censoring effect where a test result informs the clinician of a possible complication and the clinician takes appropriate and successful action to avoid the complication. Then the pattern for complication will be present, but not the complication itself, which leads to effectively mislabeled data, known as confounding medical interventions [159]. In our case this is unlikely to be a large issue since there is little information to act on to avoid a SSI such that most cases are likely to be correctly coded.

Using ICD10 and NCSP codes to phenotype a cohort there is a significant chance of miscoding

leading to labeling errors. However, there is a far greater chance of false negatives (i.e., missing coding) than false positives. In this case, the positive class will be correct while the negative class may contain erroneous labels. When generating the cohort by matching we alleviate this since a minority of patients get SSIs, and additionally we check for the Norwegian equivalent of the word "infection" in the post-operative notes, which would almost surely appear if the patient actually got an SSI.

In the literature several approaches for reducing SSI have recently been described. One of the most popular existing risk models for SSI is the National Nosocomial Infection Surveillance Basic SSI Risk Index [160]. Also, recently, a logistic regression model for predicting SSIs was developed by van Walraven et al. [161]. A more data driven approach has been used by Gbegon et al. predicting SSIs in real time within 30 days of the operation [162]. However, these studies rely on clinical data, demographic and other information but does not take blood tests into account. There exists validated risk assessment tools for post-operative complications, including the Surgical AGPAR Score [163] and the POSSUM score [164]. Both of them assess the immediate post-operative risk based on a number of variables. The American College of Surgeons' NSQIP risk calculator was developed as a preoperative risk stratification tool [165].

We have shown that our model has a potential for real time prediction and identification of patients at risk for developing SSI. This can give decision support to clinicians, and treatment plans can be adjusted taking into account the identified increased risk.

Knowledge Management in Electronic Health Record

5.1 Introduction

Health care providers require rapid and reliable decision-making processes in patient diagnosis, treatment, and follow-up. In recent decades, not only the importance of clinical decision support (CDS) systems has been widely studied [166, 167, 168], but also it is undergoing a change from traditional medical approaches based on clinicians' experiences to innovate methods based of signal and image processing for supporting clinical decisions. On the other hand, medical informatics provides a large variety of resources to healthcare community to improve many issues of their clinical daily practice [169]. In this setting, EHR can be very useful to provide access to the vast amount of clinical information and to share data among heterogenous Hospital Information Systems (HIS) [170]. However, the ability to exchange data and to understand clinical information from EHR with independence on the system (semantic interoperability) is a major challenge in this field, specially in public health systems [171].

The use of standards aims to allow the interoperability among different systems, in order to provide to citizens and professionals with the access to the same clinical information anywhere. The definition of clear and standardized connections among the current scientific knowledge, its availability for the care community, and the actual patient databases, is becoming a fundamental need for the clinical practice. In this scenario, ontologies and formal definition of clinical concepts (archetypes) can be very useful to provide a structured access to the vast amount of information in EHR, enabling the systems interoperability and the access to heterogenous sources of information [172]. First, they are useful to formalize the design of a model coping with connections among bio-signals, their representation, and the underlying anatomical,

electrophysiological, and clinical concepts. Second, they can offer advanced interoperability capabilities and foster the sharing of scientific and clinical information into a new level. And third, medical and technological knowledge can be seamless translated into decision-support tools in EHR.

On the other hand, Cardiovascular Risk Stratification (CVRS) constitutes a patient classification technique widely used in clinical practice, allowing cardiologist to focus resources on patients with higher cardiac morbidity and mortality risk. This patient classification has impact not only on the patient diagnose and treatment decision-making, but also on cost analysis, billing/funding, and clinical quality assessment. Last decade, many clinical, scientific and technical research has been devoted to CVRS [173], and many risk estimators, based on complex signal processing techniques, have been proposed to compute a set of indices from the electrocardiogram (ECG) signal included in EHR. Standardization of the CVRS domain is a complex and long-term task, not only due to the complexity of the domain itself, but also to the enormous variety of signal processing techniques proposed in the literature to calculate these indices.

In this context, SNOMED-CT (Systematized Nomenclature of Medicine - Clinical Terms) is the most comprehensive and precise clinical health terminology in the world, and it is accepted as a common global language for health terms. Most of current EHRs are adopting SNOMED-CT as their standard for the electronic exchange of clinical information [174] in the health systems of different countries [175, 176, 177, 178].

On the other hand, the goal of CEN/ISO EN13606 standard is to achieve the semantic interoperability in the EHR following a Dual Model architecture. The main advantage of the Dual Model is that knowledge is upgraded when it changes, whereas the Reference Model (information) remains unaltered. Archetypes are formal definitions of clinical concepts in the form of structured and constrained combinations of the entities of a Reference Model, providing a semantic meaning to a Reference Model structure [179, 180, 181].

Ontologies and archetypes can be used for developing a CVRS standardization framework, in close connection to the EHR, and including relevant cardiac signal processing techniques. Thus, given the generality and vastness of the CVRS domain, the Heart Rate Turbulence (HRT) is studied as indicator of CVRS, since it represents a very well established domain, with concise guidelines and clear procedures to obtain cardiac indices [173].

5.1.1 Heart Rate Turbulence

The HRT evaluation has been found to be one of the most promising noninvasive predictors in CVRS after acute myocardial infarction [173]. The term HRT describes the phenomenon of short-term fluctuation in the heart sinus cycle length over about 20 beats following a Ventricular

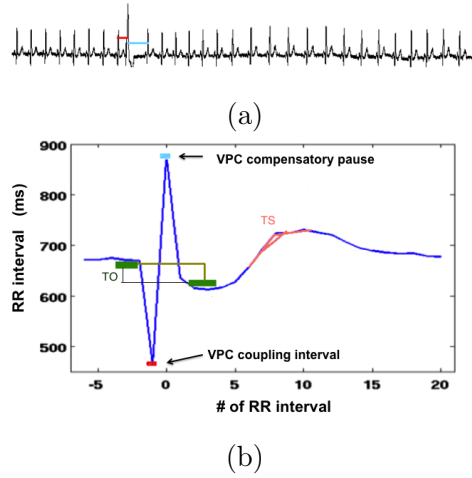


Figure 5.1: *HRT analysis and parameters extraction: (a) ECG recording with identification of VPC (the coupling interval of the VPC in red) and a posterior CP (in blue); (b) HRT tachogram.*

Premature Complex (VPC). The HRT is usually assessed from the ECG monitoring for 24 hours (24-h Holter recordings) in which the VPCs are identified and one tachogram (representation of the RR intervals versus the number of RR interval) per VPC is built. For the HRT analysis, the tachogram is built by taking 5 sinus RR intervals, the VPC coupling interval, the compensatory pause (CP, at the zero value of the abscissa-axis in the tachogram), and the subsequent 15 to 20 sinus RR interval, in this same order. An example of ECG recording, the VPC, and its corresponding HRT tachogram are shown in Fig. 5.1.

Conventionally, the HRT is quantified by two parameters, namely, the turbulence onset (TO) and the turbulence slope (TS). TO reflects the amount of sinus acceleration after a VPC, and it is calculated as the following relative difference:

$$TO = \frac{(RR_1 + RR_2) - (RR_{-2} + RR_{-1})}{(RR_{-2} + RR_{-1})} \times 100 [\%] \quad (5.1)$$

where RR_1 and RR_2 are the two RR intervals immediately preceding (following) the VPC coupling. TS measures the rate of sinus deceleration following sinus acceleration, and it is defined as the maximum positive regression slope assessed over any five consecutive sinus RR intervals within the first 15 or 20 sinus RR interval after the VPC [173].

From the obtained set of VPC tachograms during 24h, the HRT analysis can be conducted in two ways: (1) estimating parameters TO and TS from each individual VPC tachogram; and (2) averaging all individual tachograms and then estimating the HRT parameters from the averaged tachogram, according to the guidelines [173]. Note that the averaging process in the VPC-tachogram calculation could be masking physiological effects, in addition to the expected denoising. Accordingly, several advanced signal processing procedures, including SVM nonlinear regression, have been proposed for denoising individual VPC-tachograms [182]. These

methods have proven to be effective for denoising from a technical and physiological point of view, however, their impact on the CVRS capabilities has not been benchmarked with the averaged VPC-tachogram, and hence it remains an open research issue.

The HRT quantification has also been tackled with other different approaches from the signal processing point of view, such as turbulence timing, turbulence jump, TS correlation coefficient [183], or statistical detectors based on integral pulse frequency modulation model [184, 185]. On the other hand, authors in [173] reported and summarized the influence of several clinical characteristics on TO and TS parameters. As an example, left ventricular ejection fraction (LVEF) influences significantly in HRT. Furthermore, HRT is abolished at vagal blockade with atropine, and is reduced at high heart rate. Angiotension-converting receptor blockers have shown to increase both TO and TS parameters, but HRT is unaffected by beta-blockade. All these clinical characteristics have been included and represented as concepts in the proposed HRT ontology, described in detail in Sec. 5.2.1.

Absence of HRT is a noninvasive predictor of cardiac mortality following myocardial infarction. However, in other non-ischemic cardiac pathologies, HRT can remain without modification, or can exhibit demonstrable reduction but without proven value for risk stratification purposes. Watanabe and Schmidt examined in [186] whether HRT is a suitable risk predictor for mortality on patients with other non-ischemic cardiac pathologies. To cite some examples, HRT has not been determined to be a risk predictor in patients with Chagas, whereas in patients with dilated cardiomyopathy, it was significantly reduced. In contrast, HRT was not shown to be a risk predictor in hypertrophic cardiomyopathy patients. Finally, a study of 50 patients of Congestive Heart Failure (CHF) found that TS was a good predictor of rehospitalization and death.

5.1.2 The Conceptual Model of SNOMED-CT

A representation of biomedicine domains is currently being provided by the increasing use of different ontologies. For example, Unified Medical Language System (UMLS) is a repository of biomedical vocabularies developed by the United State National Library of Medicine. The Foundational Model of Anatomy Ontology (FMA) represents concepts related to the phenotypic structure of the human body in a form that is interpretable by machines [187]. The GALEN ontology contains many medical concepts, though not strongly specialized on the cardiac domain [188]. Recently, SNOMED-CT has become probably the most comprehensive biomedical terminology, with a centrally standardized and maintained clinical terminology commercially available [178]. Previous works on ECG ontologies ([189, 190, 191, 192]) aimed to give a principled approach to the cardiac domain, but their standardization with EHR is not yet warranted and they are not oriented to CVRS. The controlled vocabulary of SNOMED-CT was used in [193] to

create a patient profile ontology facilitating the semantic interoperability and the contribution of new knowledge in an heterogeneous domain. SNOMED-CT has also been used in [194] to create an ontology for the lung domain.

SNOMED-CT is an standardized and multilingual vocabulary of clinical terminology, resulting from the merging of SNOMED Reference Terminology (SNOMED RT) and Clinical Terms Version 3. Currently the not-for-profit *International Health Terminology Standards Development Organization* (IHTSDO) maintains the SNOMED-CT technical design, its core content, and the related technical documentation. Members of IHTSDO can be either agencies of national government, or other bodies endorsed by an appropriate national governments authority within the country it represents (19 countries are current members). IHTSDO distributes two releases of SNOMED-CT per year (January and July) [178].

SNOMED-CT probably represents the most complete classification for clinical use, being the reference to terminologies for different health professionals. It consists of a structured collection of health care terms, which are attached to concept codes with multiple definitions per code. SNOMED-CT is composed of concepts, descriptions (terms) and relationships, as well as other components (including extensions, reference sets, cross maps, and historical tables) [178]. A *concept* is a clinical meaning identified by a unique identifier (*ConceptID*) that never changes. Concept attributes can be used to create a new relationship among concepts. Some attribute examples are “associated with”, “severity” or “has interpretation”. The January version of 2012 includes more than 295.000 active concepts, with formal logic-based definitions organized into 19 top-level hierarchies (axes) representing body structure, clinical findings, geographic location, pharmaceutical or biological products. *Descriptions* (or terms) are the phrases used to name a concept, hence identifying a description with a unique *DescriptionID*. The January 2012 release contains more than 769.000 active English-language descriptions. Concepts in SNOMED-CT are logically defined through their *relationships*. Each active concept has at least one *is a* relationship to a super type concept (except for the SNOMED-CT root concept). It is important to note that *is a* relationships are the basis of SNOMED-CT hierarchies. There can be multi-hierarchies when a concept has more than one *is a* relationship. The January 2012 release provides more than 837.000 logically defined relationships (from a total of 1.444.673).

The SNOMED-CT **conceptual model** is used to specify logical definitions of concepts. It is based on a combination of formal logic and a set of rules determining the permitted attributes and values.

Although SNOMED-CT includes more than 295.000 active concepts, this impressive number can be not enough for representing many clinical expert domains, and for this reason, local and national *extensions* can be created. This way, contents (such as subsets of concepts) may be locally delivered by an specific clinical expert group, with the possibility to be moved to a

national extension and then to the international core.

SNOMED-CT is distributed as a set of tab-delimited text files that can be imported into a relational database. Owing to the huge number of concepts and relationships, several software tools have been developed to browse the whole terminology. CliniClue [195] is the browser used in this work.

5.1.3 CEN/ISO EN13606 standard

According to Institute of Electrical and Electronics Engineers (IEEE) interoperability is defined as the ability of two or more systems or components to exchange information and to use the information that has been exchanged. Two types of interoperability can be defined, namely, syntactic and semantic interoperability [179]. The first one refers to the capacity of communicating and exchanging data among two or more systems. Towards that end, specified data formats and communication protocols such as XML or SQL are required. This type of the interoperability is the base for any attempts of further interoperability. On the other hand, semantic interoperability is the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results as defined by the end users of both systems.

In healthcare domain, semantic interoperability is even a more important and difficult task, specially due to both scientific and technological field are changed, making necessary the development of new methodologies of information management. Among them, the dual model approach is the most promising approach. CEN/ISO EN13606 standard is to achieve the semantic interoperability in the EHR following a Dual Model architecture. The main advantage of the Dual Model is that knowledge (archetypes) is upgraded when it changes, whereas the Reference Model (information) remains unaltered.

A Reference Model is an object oriented model which compromises a small set of classes that define the generic building blocks to construct EHRs. It is used to represent the generic and stable properties of health record information [179]. The elements of the Reference Model are the following [179]: (1) a set of primitive types; (2) a set of classes that define the building blocks of EHRs, being any annotation in an EHR an instance of one of these classes (called entities). Specifically, the EN13606 standard defines six types of entities, namely, folder, composition, section, entry, cluster and element; and (3) a set of auxiliary classes that describe the context information such as the versioning to be attached to an EHR annotation.

On the other hand, archetypes are formal definitions of clinical concepts in the form of structured and constrained combinations of the entities of a Reference Model, providing a semantic meaning to a Reference Model structure [179, 180, 181]. They represent a specific clinical concept, such as blood pressure measurement. Archetypes are composed of three main

sections: header, definition and ontology [181]. The first one contains metadata such as authoring information. In the definition section, the clinical concepts is described by containing the Reference Model entities on: the range of attributes of primitive types; the existence of attributes, i.e. whether a value is mandatory for the attribute in run time data or the cardinality of attributes, i.e. whether the attribute is multi-valuate or not, among others [181]. The last section, ontology, the the entities defined in the previous section are described and bound to terminologies. Note that archetypes should be define by health domain experts.

Archetype Definition Language (ADL) is a formal language to express standard archetypes for any reference model [196, 197] using two main syntaxes: Data definition syntax (dADL) and constraint definition syntaxes (cADL). The basic structure of an ADL file consists of three main sections: (1) header, which includes specialized data and metadata about the archetype such as author, version, and status; (2) body, which contains the main formal definition of the archetype and the constraints created from the Reference Model; and (3) terminology section, which contains the ontology, the term definitions, and the map from archetype nodes to standard terminology concepts (binding). The body includes the structure and constraints of the clinical concept defined by the archetype.

In this Thesis, we focus on the standardization of the CVRS based on HRT to achieve the semantic interoperability among different EHRs. Towards that end, a HRT ontology and various archetypes were built. Furthermore, a web prototype, based on the ontology and archetypes, was created to overcome the technical limitations found when working with different HIS.

5.2 Application 5.1: Ontology for Clinical Decision Support in EHR

This work has been devoted to create a HRT ontology yielding a well defined representation of the HRT domain in the EHR context. Two possible uses for the HRT ontology in the EHR context are next addressed: (1) clinical HRT recordings, often maintained by scientific societies, with simple access to conventional and new EHR fields by means of SNOMED-CT periodic updates; (2) retrospective and/or prospective scientific research studies, involving the patient information in EHR, including further patient detailed information and enabling signal processing capabilities. Recall that, since HRT indices are not fully established, their current use is not feasible in the medical routine yet. In this sense, this work aims to be a contribution towards its extensive use and exploitation in the clinical practice.

Two use studies are next presented. The first one corresponds to the practical implementation of the HRT ontology for its support to regional and national recordings in medical scientific societies. It can be considered as a prototype, and has been implemented in the HIS of the

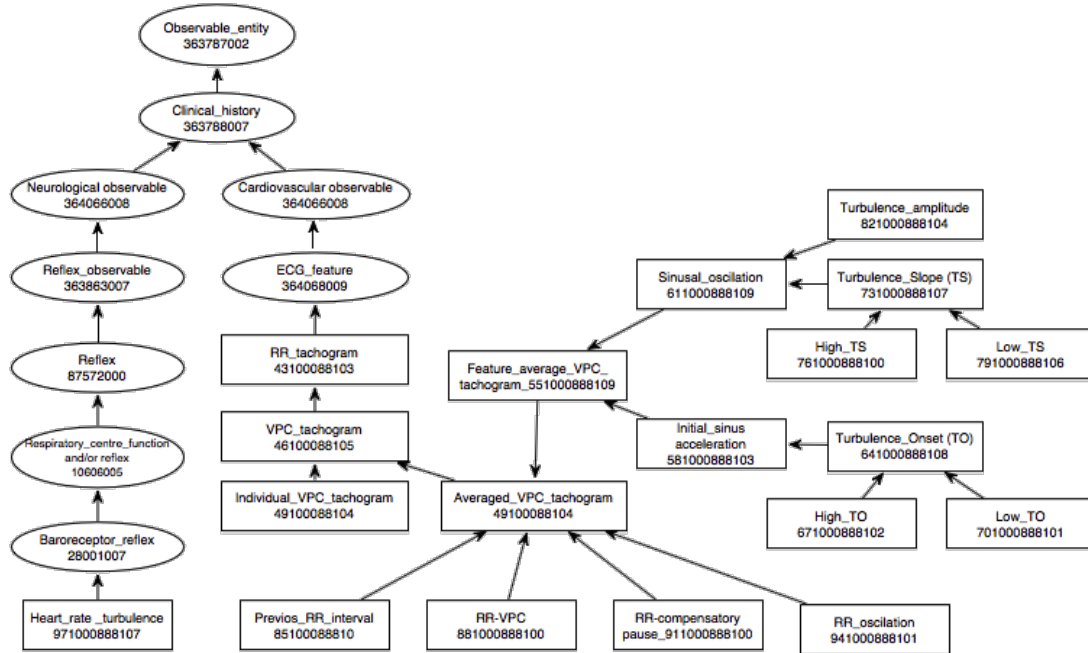


Figure 5.2: Schema of some concepts and the corresponding Concept Id for the proposed HRT ontology. Circles (rectangles) indicate original (extended) concepts. Arrow direction indicates a relation is-a between origin and destination.

Table 5.1: Some SNOMED-CT Object Properties in HRT ontology. Local extended concepts are in italic.

Concept 1	Object Property	Concept 2
Normal heart rate_78663003	Interprets_363714003	Cardiac conducting system structure_24964005
Disorder of cardiac function_105981003	Finding site_363698007	Heart structure_80891009
Disorder of cardiac ventricle_415991003	Finding site_363698007	Cardiac ventricular structure_21814001
24 hour ECG_252417001	Method_260686004	Monitoring-action_360152008
24 hour ECG_252417001	Using device_424226004	Electrocardiographic monitor and recorder_74108008
<i>HRT_971000888107</i>	<i>is a measurement of_161000888104</i>	Baroreceptor_reflex_28001007
<i>TO_641000888108</i>	<i>is influenced by_191000888105</i>	Left_ventricular_ejection_fraction_250908004
<i>TS_731000888107</i>	<i>is influenced by_191000888105</i>	Left_ventricular_ejection_fraction_250908004
<i>TS_731000888107</i>	<i>is calculated in_101000888100</i>	<i>Averaged tachogram_491000888104</i>

University Hospital of Fuenlabrada (Madrid, Spain). The second use study is a simple application example for medical data support involving EHR and signal processing techniques. Note that the number of hospital centers that can be joined to this research study can increase in relation to just using conventional follow-up mechanisms (such as handwritten documental support or web technology server support). Both studies are based on the HRT ontology, thus, its construction is first described.

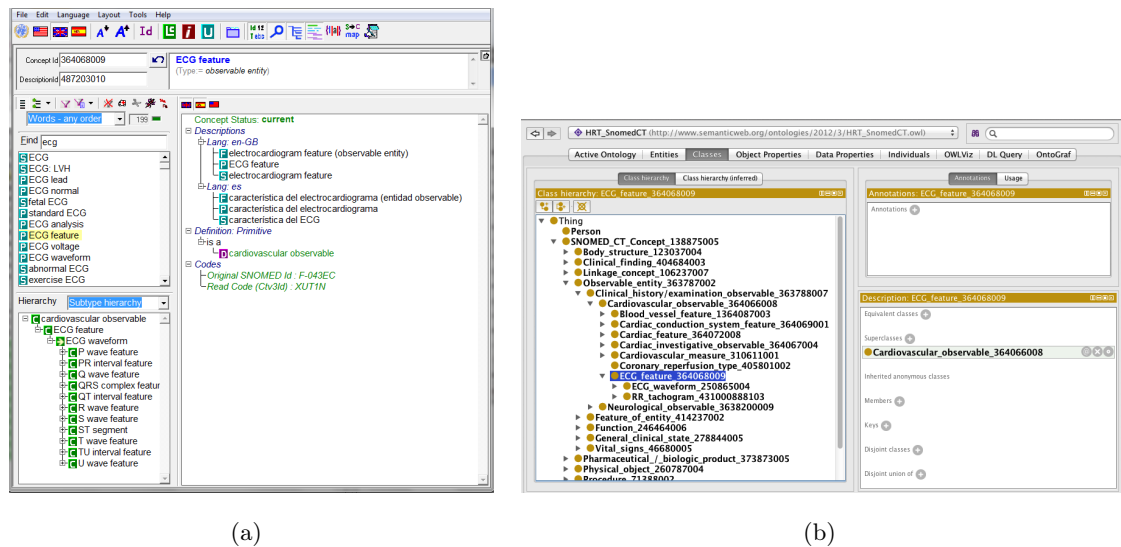


Figure 5.3: Software tools for ontology design support and SNOMED-CT browser: (a) CliniClue Xplore screenshot, with ECG feature concept and its hierarchy, status and identifier; (b) Protégé screenshot remarking one concept and its corresponding SNOMED-CT concept identifier.

5.2.1 Ontology based on SNOMED-CT

Several languages have been proposed for building ontologies. Both Resource Description Framework (RDF) and Web Ontology Language (OWL) are based on the Extensible Markup Language (XML), as proposed by the World Wide Web Consortium (W3C) [198, 199, 200]. OWL has been chosen in this work because it is widely-used as standard for representing and sharing knowledge in the semantic web context [201]. Furthermore, it can provide additional vocabulary with formal semantic, hence yielding a greater machine interpretability than that supported by XML or RDF.

The construction of the HRT ontology using OWL was based on the methodology proposed in [202], with the following steps:

(1) *Determine the domain.* The goal in this work has been the HRT domain (previously defined in Section 5.1.1).

(2) *Enumerate the relevant concepts.* First, a set of clinical, anatomical, electrophysiological, and pharmacological features, were identified in order to provide an accurate representation of the HRT domain. Most of the clinical concepts were based on existing SNOMED-CT concepts. However, a subset of 19 concepts (over 308) could not be taken from SNOMED-CT and were extended using the *Fuenlabrada Hospital* namespace. Figure 5.2 shows a schema of some concepts and their *is-a* relationships, where the 19 extended concepts are in rectangle.

(3) *Define the concepts* and their hierarchy (down-top development process). HRT concepts directly mapped on the SNOMED-CT terminology followed the SNOMED-CT hierarchy. For the

19 extended concepts, the SNOMED-CT axis associated to each concept was first decided, e.g., in a *body structure*, in a *clinical finding*, or in an *observable entity*. Then, the concept was assigned to its parent class. The SNOMED-CT January 2012 release was systematically examined using the CliniClue terminology browser [195] to determine whether the defined concepts were represented in the terminology. Figure 5.3(a) shows an example of a defined concept (ECG feature), its hierarchy, status and identification (Concept ID). In Fig. 5.3(b), a screenshot of Protégé (an open source ontology editor) shows the ECG feature concept in our HRT ontology hierarchy. The concept ID was used to extend the term of the ECG feature concept, what could be used in future to standardize the ontology concepts, in such a way that they can be used in any EHR. A public available version of these concepts and their relationships *is-a* can be found at [203].

(4) *Define the properties of the concepts.* There are two properties. Data Properties define features associated with one concept, whereas Object Properties define relationships between two concepts. Table 5.1 presents some concepts and Object properties used in the HRT domain. Extended concepts and extended properties are shown in italic. Note that extended concepts require new properties to relate them to other concepts.

(5) *Use the inference mechanism,* in this work, to get a risk stratification criterion for patients with CHF. Note that the logic description implemented by OWL does not offer inference capabilities. Semantic Web Rule Language (SWRL), the rule language to infer new knowledge [204], has been considered in this work for this purpose. SWRL extends the set of OWL axioms to include Horn-like rules [205] for enriching the ontology with logical rules descriptions to step up the knowledge discovered by the concepts, by their relations, and by the mining process (when it is considered).

5.2.2 Ontology Prototype in EHR

Scientific medical societies often maintain regional (or national) recordings for specific and relevant aspects of their field, for instance, the implantable cardioverter defibrillator recording from cardiology societies [206]. These recordings include basic information about the patient data and additional information about the specific subject, mostly compiled by accessing to EHR, handwritten additional documentation, and in best cases, web technology support. The previously described HRT ontology has been used to implement a prototype providing with advanced support to a HRT recording. The HRT ontology is being tested in the HIS of University Hospital of Fuenlabrada (SeleneTM system from the Siemens company). In this prototype, the most relevant clinical variables in the HRT domain have been gotten into three groups: (1) Patient data, such as age, smoking, or alcohol; (2) Domain concepts, such as structural heart disease, other heart disease, heart failure, history of arrhythmic episodes (atrial fibrillation, ventricular tachycardia, ventricular fibrillation, sinus dysfunction), other arrhythmic

HRT Cardiovascular Risk	
Age (años)	37
Tobacco	
Alcohol	< 110 gr/week
Diabetes Mellitus	No
Glycated hemoglobin (%)	20.0
Arterial Hypertension	Primary
Drugs	Calcium blocking agent
Other drugs	
Structural heart disease	Dilated cardiomyopathy
Other heart disease	
Heart failure (disease)	Class II
HISTORY OF ARRHYTHMIC EPISODES	
Atrial fibrillation	<input type="checkbox"/>
VT - VF	<input checked="" type="checkbox"/>
Sinus dysfunction	<input type="checkbox"/>
Other arrhythmic episodes	
Non cardiac diseases	Renal failure
Other non cardiac diseases	
TO value	0.0
TS value	3.0
HRT evaluation	Category 1

Figure 5.4: *Developed and implemented prototype of HRT ontology in EHR, for supporting regional and national recordings from scientific societies in the context of HRT follow-up with basic data.*

episodes, non cardiac diseases, other non cardiac diseases, and TO and TS values; (3) other relevant concepts for decision making, such as diabetes mellitus, glycated hemoglobin, arterial hypertension, or drugs. The prototype infers the risk in terms of the TO and TS values from 24h Holter data.

Note that a key issue is the calculation of TO and TS from signal processing techniques considering 24h Holter recordings. Though this calculation can be manual and offline, it is possible to automate this process by accessing to a specialized remote server enabling the use of signal processing techniques on the Holter recordings.

The information compiled by the prototype can be used to create an HRT recording to give support to the health centers with reduced cost and effort, undoubtedly increasing the knowledge on actual HRT incidence in connection with basic data in EHR.

Figure 5.4 shows a screenshot of the prototype. Though it has been created by a reduced expert committee, remark here that it should be validated by the corresponding scientific society to reach a consensus on other factors influencing the HRT and their inclusion in a prototype

Table 5.2: Cardiovascular risk stratification results.

Avg. tach.			Indiv. tach.			Indiv. tach. with SVM			Total tach.
C0	C1	C2	C0(%)	C1(%)	C2(%)	C0(%)	C1(%)	C2(%)	
-	-	x	13.5	46.0	40.5	0.0	10.8	89.2	37
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	1
-	x	-	65.0	35.0	0.0	65.0	35.0	0.0	40
x	-	-	90.7	9.3	0.0	98.8	1.2	0.0	86
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	1
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	1
x	-	-	89.0	0.0	11.0	77.8	22.2	0.0	9
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	1
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	1
x	-	-	50.0	50.0	0.0	100.0	0.0	0.0	2
x	-	-	69.6	30.4	0.0	87.0	13.0	0.0	23
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	1
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	1
x	-	-	100.0	0.0	0.0	100.0	0.0	0.0	2
-	x	-	0.0	100.0	0.0	0.0	100.0	0.0	1
x	-	-	82.1	17.9	0.0	94.6	5.4	0.0	56
-	x	-	0.0	100.0	0.0	0.0	100.0	0.0	1
-	x	-	0.0	100.0	0.0	0.0	100.0	0.0	1
-	x	-	0.0	100.0	0.0	0.0	100.0	0.0	1
-	x	-	50.0	50.0	0.0	0.0	100.0	0.0	2
-	x	-	0.0	75.0	25.0	0.0	25.0	75.0	4
-	-	x	0.0	0.0	100.0	0.0	0.0	100.0	1
x	-	-	100.0	0.0	0.0	66.7	33.3	0.00	3
x	-	-	100.0	0.0	0.0	100.0	0.0	0.00	1
-	x	-	0.0	100.0	0.0	0.0	100.0	0.00	1
-	x	-	50.0	50.0	0.0	0.0	100.0	0.00	2
-	x	-	60.9	34.8	4.3	43.5	56.5	0.00	69

extension. For instance, factors such as cardiopathy, drugs and non-cardiac illness, which can affect HRT, can be present simultaneously; a patient can have ischemic cardiopathy as well as valvular cardiopathy or can have beta-blockers and amiodarone at the same time. Hence, determining the need for multiple choice or main description has to be established. Also, relevant cardiological information for risk stratification which can be readily obtained from the EHR has to be taken into account, such as ejection fraction.

5.2.3 HRT Clinical Decision Support

As a second use case, the proposed HRT ontology was used to provide support to a study with ECG recordings. The study has a multidisciplinary clinical and technical scope, and aimed to compare both conventional and emergent signal processing methods for calculating HRT parameters described in Section 5.1.1. The goal is to analyze the effect of applying different signal processing techniques for the HRT indices calculation. This analysis is limited to the comparison of the TO and TS values (HRT indices) obtained with three procedures involving the tachogram, namely: (1) conventional averaged-tachogram; (2) individual tachograms without denoising; and (3) individual tachograms with SVM denoising for each tachogram (see [182] for signal processing details).

We should take into account that, in most clinical studies, values of $TO < 0\%$ and $TS > 2.5$ ms/RR-intervals are considered as normal after acute myocardial infarction and monitoring of disease progression in CHF [173]. However, the cut-off values used for risk stratification of patients suffering other heart diseases are not clearly defined. In healthy subjects, up to four studies have reported that averaged TO ranged from -2.7% to -2.3% and averaged TS ranged from 11.0 to 19.2 ms/RR-intervals [173]. A cut-off value of 3 ms/RR-intervals for TS was proposed as the optimal stratification value for patients with CHF [207]. In the MUSIC trial, a prospective multicenter longitudinal study was designed to assess risk predictors in patients with CHF, identifying high-risk quartiles as $TS \leq 1.27$ ms/RR-intervals and $TO \geq 0.25\%$.

In this work, a database of 24h Holter recordings from CHF patients has been collected in the Arrhythmia Unit of University Hospital Virgen de la Arrixaca (Murcia, Spain). RR-tachograms were analyzed to identify reliable individual VPC tachograms, according to the criteria proposed in [173]. Up to 27 of 60 recordings were useful for this analysis (14 recordings with only one individual VPC tachogram). According to guidelines [173], cut-off values of 0% and 2.5 ms/RR-intervals for TO and TS, respectively, were used to classify HRT indices into three categories. First, C0 when both TO and TS are normal (i.e. $TO < 0$ and $TS > 2.5$); then, C1 when either TO or TS is abnormal (i.e. $TO > 0$ or $TS < 2.5$); and finally C2 when both TO and TS are abnormal (i.e. $TO > 0$ and $TS < 2.5$). OWL can deal with classification problems using description logic, but many applications for risk stratification (as it was in this work) require ontologies and rules together, i.e., SWRL [208].

Table 5.2 shows the 27 patients of our database and their corresponding classification according to the three previously procedures: conventional averaged VPC tachogram criterion (first column, *Avg. tach.*), individual VPC tachograms without denoising (second column, *Indiv. tach.*) and individual VPC tachograms with SVM denoising (third column, *Indiv. tach. with SVM*). Total number of individual tachograms is presented in last column (*Total. tach.*). The category assignment is indicated with a cross for the first procedure. For the two remaining





Concept	Description	Constraints	Values
 Patient Date	Patient summary	Cluster 0..1	
 Age	Patient age	Count 0..1	0..120
 Gender	Patient gender	Text 0..1	Internal, Male, Female, Unknown
 Tobacco	Tobacco usage	Cluster 0..1	

Figure 5.5: Example of concepts, descriptions, constraints and values from the Patient Summary archetype in HTLM format.

procedures, Table 5.2 shows the percentage of individual tachograms assigned to each category (boldface for the majority percentage); the patient is assigned to the majority category.

Parameters obtained by conventional averaged tachogram classified 15, 10, and 2 patients on C0, C1, and C2, respectively. For comparison purposes, patient classification according to raw individual VPC tachograms, was 16, 7, and 1, with no majority percentage for 3 patients. The procedure based on individual SVM filtering yielded 16, 8, and 3 patients. Note that there are 3 patients (11%) with more than 30 tachograms which are differently classified when using conventional and individual raw tachograms. Comparison of conventional and SVM denoising procedures shows that just 2 patients (7%) are classified in different categories, one patient with 40 beats is assigned to C0 (underestimation regarding conventional procedure) and another one with 4 beats is assigned to C2 (overestimation regarding conventional procedure).

From the previous analysis with a reduced number of patients, it can be said that the class assignment is dependent on the signal processing algorithm previous to TO and TS calculations. The most adequate procedure could be assessed by increasing the number of patients. This could be achieved with affordable technical and medical effort by means of the inclusion of the HRT ontology in the EHR. Remark here that the use of a remote server for indices calculation using signal processing techniques has shown to be convenient (information about the project is available in spanish at: <http://vpredict.org/formacion>; research results will be presented in a dedicated work).

5.3 Application 5.2: From Archetypes to EHR Web Prototype

In the previous application, an ontology was built based on the conceptual model of SNOMED-CT for CVRS using ECG-derived indices. The ontology was focused on the current knowledge of HRT, since it represents a low complexity model domain (see [209]). As previously described, two use studies were proposed therein. The first one corresponded to the practical development of a clinical form based on the HRT ontology, which was implemented in the HIS of University Hospital of Fuenlabrada (Madrid, Spain). The second one was a simple application example for CDS involving the EHR and signal processing techniques. Two main drawbacks

were found in these practical implementations. On the one hand, the implementation of the HRT form in different HIS was very difficult to reach, due to changes in commercial systems requiring a political consensus. On the other hand, the HRT form based on the ontology did not achieve semantic interoperability in the EHR because it was created with no clinical standard to represent the relevant health information.

To overcome these drawbacks presented previously, the aim of this work was to create a rigorous and stable information architecture for communicating some parts (or all) of the EHR of a single subject of care (patient) among heterogenous HIS, by using a web prototype to account for the technical requirements of different systems. To achieve this goal, a CVRS archetype following the CEN/ISO EN13606 standard to achieve the interoperability among heterogenous HIS is proposed in this work. A simple and helpful way of binding archetypes to the HRT ontology is also addressed.

Apart from that, a threefold web prototype to get semantic interoperability among heterogenous EHR systems was developed in this work. First, a web prototype, called *HRT Archetype Proto* was built from clinical archetypes, so that their advantages in the EHR are also maintained in the prototype. It is proposed an HRT archetype transformation which allowed importing the nodes of the archetypes as fields in a *MySQL* HRT database. Data generated by the prototype were saved in this database. Second, a server-based ontology system has been incorporated into the prototype for binding the nodes of the archetype to the HRT ontology. Third, the EHR data were exported in *xml* files, which allowed their sharing among heterogenous systems.

The web prototype supports: (1) the use of clinical standards for CDS; and (2) the development of a structured database to scientifically assess and improve the knowledge of the HRT domain for allowing the wider and subsequent CVRS domain expansion.

5.3.1 HRT Archetype Prototype

In this work, the standardization of different domains using the CEN/ISO EN13606 standard as recommended in [210] is proposed. Specifically, the recommendation suggests that semantic interoperability is an essential factor in achieving the benefits of EHR to improve the quality and safety of patient care, public health, clinical research, and health service management. The Commission in [210] encourages the use of standards to represent the relevant health information for a particular application using data structures (such as archetype and templates), terminology systems and ontologies.

Archetype Editor software [211] was used to build an HRT archetype and a Patient Summary archetype. A group of clinicians from different hospitals agreed that two archetypes were needed to separate the summary of the patient data from data related to HRT CVRS. These two

Atrial Fibrillation:	True False
VT / VF:	True False
Sinus Dysfuntion:	True False
Other Arrhythmic Episodes :	Free text
TO Value:	0,00
TS Value:	0,00

Figure 5.6: Screenshot of the HRT archetype nodes.

archetypes were used to model and improve the knowledge of these domains. Furthermore, clinicians defined all the nodes and constraints of both archetypes, achieving the best knowledge representation from a clinical viewpoint. An example of a set of concepts, definitions, constraints and values from the Patient Summary archetype are shown in Fig. 5.5. This archetype records health information potentially useful for HRT CVRS. It also includes factors such as hypertension, gender, or previous heart diseases. Free text is only allowed in fields “other drugs” and “other heart disease”, allowing clinicians to include additional information.

On the other hand, the HRT archetype compiles the information of 24h Holter recordings, in order to infer the CVRS in terms of HRT indices (see Fig. 5.6). CVRS was inferred using TS and TO values, considering the cut-off values and categories described previously. Since the clinical evidence of other factors influencing the HRT has not been clearly validated yet, they were not considered to infer CVRS. However, they were recorded for future follow-up, and for accounting for this kind of clinical information in the web prototype.

In order to obtain semantic interoperability, knowledge and information were separated by following a twofold schema. On the one hand, knowledge is provided by clinical experts by defining the domain elements directly using an archetype editor software, which was *Archetype Editor Ocean Informatics* in this implementation. On the other hand, information is managed and supported by the web prototype *HRT Archetype Proto*. Figure 5.7 (a) shows the followed approach: first, archetypes were built; second, the *ADL* files obtained into a *csv* file were exported to extract the information related to nodes of the archetypes and their constrains; and third, this *csv* file was used to generate a clinical standard table in a *MySQL* database, hence allowing the development of *HRT Archetype Proto*.

A manual process was followed to incorporate in the web prototype all the nodes and constraints extracted from the archetypes. In particular, constraints such as ranges of allowed

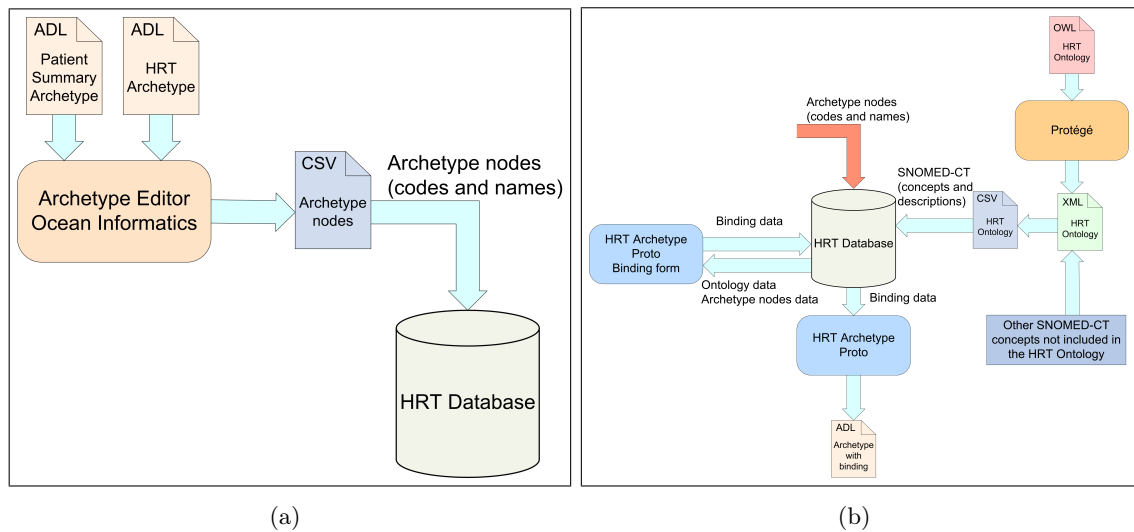


Figure 5.7: Schema for: (a) exporting the information related to the nodes of the archetypes into fields of the HRT database; (b) the server-based ontology system integrated in the HRT Archetype Proto Web prototype.

values and coded text options were considered, as shown in Fig. 5.5. *Apache*, *PHP*, *MySQL* and *phpMyAdmin* were installed under the web development environment *WampServer* to design the prototype. The web prototype not only allows management of patient data and EHR (add, edit, view, delete), but also, it provides with other functionalities: (1) it allows binding the nodes of archetypes to concepts drawn from HRT ontology; and (2) EHR exporting in order to enable sharing of clinical information.

5.3.2 Server-Based Ontology System Archetype Binding

Binding process consists on the structural relationship between the archetypes nodes and the terminology concepts. Several methods have been proposed for automatic or semi-automatic binding. For example, authors in [212] proposed an archetype editor which supports a manual or a semi-automatic binding process. However, this editor has not been updated since 2008. In [213], an automatic binding mechanism based on an information retrieval system was evaluated.

In this work a solution to help clinicians to use the previously defined ontology without requiring either a terminology server or a specific software is proposed. Specifically, a simple system to bind SNOMED-CT concepts from the HRT ontology with the nodes of the HRT archetype were built. For this aim, the schema shown in Fig. 5.7 (b) was followed. As first step, *Protégé* software tool was used to export the HRT ontology created in [209] into an *xml* file. This file was converted to *csv* format to be able to easily add other SNOMED-CT concepts useful for binding which were not included previously in the HRT ontology. The obtained *csv* file was

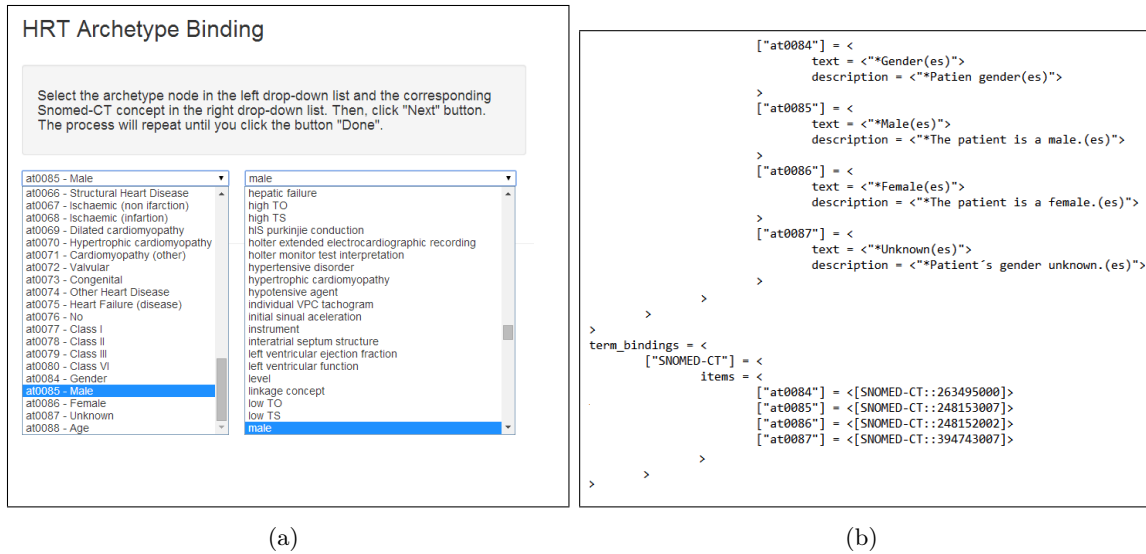


Figure 5.8: *Binding process: (a) Screenshot of the binding process integrated in the HRT Archetype Proto Web prototype; (b) Example of ADL file after the binding process.*

imported into a table in the HRT database. As a second step, the archetypes nodes (codes and names) were also imported into a table in the HRT database as explained in the previous section.

Finally, the *HRT Archetype Proto Web* prototype automatically generates two drop-down lists that allow the user to match archetypes nodes with the concepts from the ontology (see Fig. 5.8 (a)). These pairings are temporarily stored in a table in HRT database, and deleted once the entire process has finished. With the binding data stored in the HRT database, the *HRT Archetype Proto* web prototype, automatically writes into the portion corresponding to the binding code of the original *ADL* file (see Fig. 5.8 (b)), hence completing the binding process. The web prototype allows the user to download the generated *ADL* file with the binding terms, therefore, it can be now readily shared among different systems.

5.3.3 Clinical Data Export for Semantic Interoperability

The interoperability among different health care systems was pursued in this work by: (1) creating a structured data based on the standard CEN/ISO EN136066; (2) using SNOMED-CT, since it guaranteed the standardization and interoperability with emerging EHR; and (3) exporting clinical data in *xml* to be shared with fully meaning among different systems. The last process was developed according to the schema shown in Fig. 5.9 (a). First, the *Template Designer* software [214] was used to export the information of the path of each archetype node and its datatype and classtype in an *xml* file. An example of the structure of the *xml* file is shown in Fig. 5.9(b). Next, data of this template were imported into a table in the HRT database. This database contains all data generated when a new EHR extract is created using

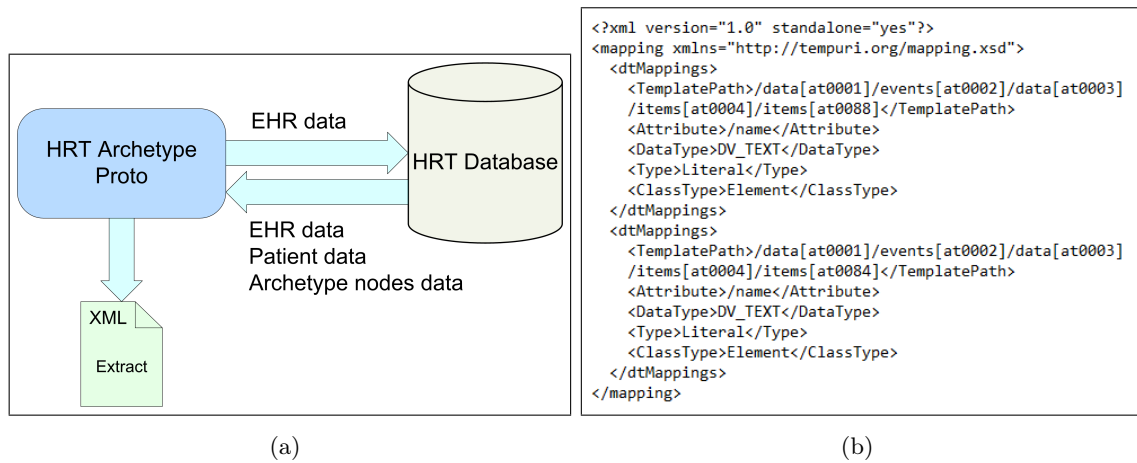


Figure 5.9: *Exporting clinical data in xml: (a) schema for semantic interoperability; (b) Extract of the xml file obtained from Template Designer software with path, datatype and dataclass information..*



Figure 5.10: *Extract of the xml EHR file generated by the HRT Archetype Proto Web prototype.*

the web prototype *HRT Archetype Proto*. Finally, the web prototype generates *xml* files in the form of EHR extracts, by combining the data from the nodes of the archetype and the EHR data entered by the clinicians, both stored in the HRT database. These extracts have the same structure and constraints as the built archetypes, so they provide the same advantages, i.e. semantic interoperability. Figure 5.10 presents an example of an *xml* file generated by the HRT web prototype.

The information compiled by the prototype can be used to create standard HRT records to give support to the health centers with reduced cost and effort, undoubtedly increasing the knowledge on actual HRT incidence in connection with basic data in the EHR.

5.4 Discussion and Conclusions

This work has addressed the proposal and building of an ontology for CVRS based on ECG-derived indices, focused on the current knowledge of HRT due to its well-defined procedures for parameters calculation and concise guidelines. Other ECG-derived indices, such as Heart Rate Variability (HRV), T-wave alternans (TWA), deceleration capacity, and many others, have also been proposed to assess CVRS [215, 216, 217], however, their knowledge domains are significantly more extensive, and consequently they will require further and specific work. The HRT ontology has been used here for putting together for the first time most of the relevant elements of HRT, standardization, ontologies, and SNOMED-CT, in the CVRS setting, considering signal processing resources as well.

The present work has also addressed the proposal and building of a web prototype, called *HRT Archetype Proto*, for achieving semantic interoperability among heterogeneous EHRs, by using ontologies and archetypes. *HRT Archetype Proto* provides the user with several novel functionalities. First, the creation and maintenance of EHR extracts from the knowledge of the HRT archetype; second, the binding process using the SNOMED-CT concepts from the HRT ontology; and third, the exporting of these extracts in *xml* files, hence allowing their sharing with fully semantic meaning among different systems, since these extracts are generated from archetypes. Interoperability with EHR has been reinforced by applying the conceptual model of SNOMED-CT, hence allowing a clear identification of the directions to put the effort in oncoming work.

Two representative use studies have been first presented. First, a prototype was implemented in the University Hospital of Fuenlabrada (Madrid, Spain), in order to support HRT recordings simple follow-up by medical societies. Second, a simple application of risk categorization has been used for yielding a straightforward comparison between conventional and emergent signal processing techniques in the HRT indices calculations, and its impact on the classification of a patient database. The long-term objective is the use of HRT ontology in particular, and CVRS in general, in the EHR context for efficient support to the clinical practice. As this is not yet feasible today, given that HRT and other ECG-derived indices for risk stratification are not established in the clinical practice, we consider our work as a first step and a significant contribution towards its use in the near future. The methodology can be readily extended to other relevant ECG-derived indices for CVRS, such as HRV or TWA. A systematic review of the

scientific and technical literature is mandatory, due to the vast amount of data with different scientific evidence that can be found today in these fields.

Overall, *HRT Archetype Proto* allows a new web system for HRT CVRS decision support based on clinical data standards and on the SNOMED-CT conceptual model. The consensus of the specific and relevant aspects of HRT domain by clinicians is a necessary step to reach and maintain the records for this field. The web prototype allows semantically interoperable HRT records to compile and exchange with fully meaning information about HRT as well as with basic patient data. In this work, only HRT indices were considered to infer CVRS, although more cardiac factors are compiling for future following-up. However, other factors which can affect HRT, such as drugs and noncardiac illness, should be recorded in different archetypes due to they represent different knowledge domains.

From a clinical point of view, CVRS has to take into account factors depending on the patient cardiopathy, being the HRT just one of them. As shown in Section IV-A, our current prototype will provide the patient risk category just in terms of TO and TS values. Additional variables in the prototype have been included because they are fundamental for better interpreting the turbulence indices. Two examples of the convenience of complementing the information provided by TO and TS for HRT characterization are the following. First, be a patient with hypertrophic cardiomyopathy and taking drugs as beta-blockers, assigned to C2 according to TO and TS values. From the literature, TO and TS have no prognostic value in this group of patients, and classification would not be used as support for clinical decision. Second, be a patient with myocardial infarction, ejection function of 35%, 2nd degree stress dyspnea and implantable defibrillator indicating border-line ischemic cardiopathy. Clinical guides indicate that risk stratification must be based on ejection function and degree of functional state of the patient. However, in border-line cases like this one, other stratification elements can be taken into account, such as HRT, HRV, TWA, presence of non-sustained ventricular tachycardia, or electrophysiological study inducibility, among others. None of them is strong enough for changing a therapy indication, but their joint consideration can help us to take the clinical decision in these cases. Note that, both for clinical and research use, variables potentially modifying the HRT indices have to be taken into account, e.g. if the patient is smoker, diabetic, or takes drugs, despite they are not stratification indices.

Conclusions and Future Work

6.1 Conclusions

This Thesis has dealt with the use of ML techniques for analyzing data from fields as diverse as promotional efficiency and healthcare. Towards that end, several contributions to ML literature, with emphasis on the FS and predictive model design stages, have been addressed. Specific procedures for each stage have been proposed with applicability to real-world tasks of different nature.

In order to tackle the general objective, ML techniques have been proposed to adjust the predictive model to data characterized by high dimensionality, sparsity, temporal dynamics, and scarcity in the number of samples. Specifically, a feature engineering approach, a smoothing regression method based on the properties of the covariance matrix, three different FS strategies based on statistical principles, and a methodology for predictive models benchmarking, have been proposed and described in this Thesis. The first specific objective was addressed by proposing a novel data-driven approach to characterize promotional efficiency at both store and chain level. The second and third specific objectives belonged to healthcare domain. The goal of the second objective was to infer new knowledge from complex heterogenous longitudinal records of patients for supporting the early detection of several complications after CRC surgery. The third specific objective consisted on opening the road towards achieve the semantic interoperability in EHR data exchange and follow-up.

From the outcomes of the research activity developed within this Dissertation, several conclusions can be drawn. The specific conclusions for each application have been described in detail in its devoted chapter. Therefore, only a summary of the main conclusions obtained for both the general and the specific objectives are next compiled.

Theoretical Fundamentals and Contributions in ML

As a general conclusion, it can be stated that the theoretical ML contributions proposed and developed in this Thesis have been implemented in several real-world applications with highly diverse nature. The most relevant contributions in Chapter 2 are next summarized:

- A nonparametric approach based on bootstrap resampling to individually characterize the dynamics of each feature.
- A statistical procedure to represent the output based on multidimensional feature spaces by simply bootstrapping the available observations. Thus, yielding the necessary statistical distribution for their characterization without losing neither temporal information nor joint feature relationships.
- A new regression method inspired by covariance properties used in GPs to obtain a smoothed version of a original random process, achieving a denoising process and an imputation approach when working with sparse data.
- Three novel FS strategies based on the weights obtained by a SVM linear classifier: (a) a simple statistical criterion based on leave-one-out; (b) an intensive-computation statistical criterion based on a bootstrap resampling approach; and (c) an advanced statistical criterion based on KECA.
- An operative benchmark methodology based on a cut-off nonparametric statistical test to characterize the model generalization.
- A simple nonparametric statistical tool, based on the paired bootstrap resampling, to allow an operative result comparison among different ML models.

A summary of the conclusions for each specific objective is next described.

ML for Promotional Decision-Making

As a general conclusion in the promotional decision-making area, it can be stated that the use of bootstrap resampling allows benchmarking statistically the performance of the different methods discussed here. However, obtained results are influenced by the promotional method characteristics, as well as by the specific nature of the product. So, overall conclusions cannot be drawn for products and categories, and it is necessary to benchmark more than one method when building promotional estimates from real-world data. Regarding to the promotional chain-level analysis, data have been first aggregated, and a set of new indicators based on bias

and scatter measurements in the input feature space have been then proposed to take into account the reliability and stability of the promotional models.

ML for Healthcare Analytics

A learning system based on ML techniques was also proposed in Chapter 4 for supporting the early detection of complications after CRC surgery. The use of structured and unstructured data from the EHR have a potential for early warning and decision support, despite the challenges related to the veracity and completeness of the data. Combining free text and the temporal structure of blood tests and vital signs for pre- and post-operative early warning dramatically improved the system accuracy. This can provide the basis for future on-line systems alerting clinicians about patients at risk for complications, so that appropriate actions can be taken.

Knowledge Management in EHR

The work presented in Chapter 5 has addressed the proposal and construction of an ontology and several archetypes for CVRS based on ECG-derived indices, focused on the current knowledge of HRT due to its well-defined procedures for parameters calculation and concise guidelines. Semantic interoperability is an essential factor in achieving the benefits of EHR to improve the quality and safety of patient care, public health, clinical research, and health service management. Towards that end, in this work a web prototype was built based on the use of standards to represent the relevant health information for a particular application, by using data structures (such as archetypes and templates), terminology systems, and ontologies.

6.2 Future Work

The analysis conducted so far constitutes a step forward into the understanding of how to modify and use ML techniques to support decision-making in diverse real-world applications. However, there is still a lot to explore and understand in both, theoretical and practical areas, hoping that the results presented in this Thesis will encourage further investigation on this and other related topics.

Theoretical Fundamentals and Contributions in ML

One of our short-term research goals from a theoretical viewpoint is to investigate further on each of the contributions proposed in this Thesis, being the following the main future lines of

research:

- *Feature engineering.* The use of a temporal (dynamic) approach to deal with highly sparse data, might lead to a deep understanding of the nature of each feature.
- *Predictive modeling.* On the basis of our proposal related to covariance kernel series, a further analysis can be done for using it in semi-supervised tasks.
- *Feature selection.* We only used a state-of-art non-linear FS method in this work, but more theoretical and experimental work should be devoted to the topic of large input spaces data. For example, the proposal of a stable nonlinear FS method based on the weights of the SVM to select relevant features to the predictive modeling. Note also that sample imputation and FS are strongly coupled problems in this kind of sparse temporal data, so it is recommendable to develop methods for their joint assessment.

The main future lines of research related to the specific objectives are next described.

ML for Promotional Decision-Making

The work presented in this topic is a starting point for future research that can be oriented towards improving the proposed method or extending the results and methods. In particular, the studies developed for milk and beer, as well as for laundry detergent, could be extended to a wider number of categories, to determine whether a higher grouping scheme, would eventually allow to validate some of the proposed conclusions on a wider scope. New categories, such as perfumery or perishable product could be also analyzed. In addition, the joint modeling of different complementary or alternative categories could incorporate further information on cross-effects in these related products. New approaches could be also explored by introducing a priori information, such as environmental or idiosyncratic variables. Finally, more advanced ML techniques could be used and evaluated for improving the predictive capabilities of nonlinear promotional models.

ML for Healthcare Analytics

In future work, more advanced natural language process tools will be incorporate to build models that may be more robust to erroneous inputs, such as misspelled or accidentally omitted words. However, most available methods are designed for English language, and not directly applicable for Norwegian or Spanish clinical language. For English, a common practice is to use the UMLS [174], which enables a consistent representation of clinical language, to which no Norwegian counterpart exists to our knowledge. These issues represent a challenge for future work. The results of this work indicate that the clinical narrative, in combination with some

structure data, can be used as the basis for a clinical decision support system. A complete system should consider all available information as outlined above, and should be designed both in collaboration with clinicians and EHR providers to build a streamlined, usable, and useful system integrated in the patient care. There are significant technological and operational challenges, but establishing robust and methodological methods is a relevant first step in the process to design such systems and integrating them in the surgical workflow.

Knowledge Management in EHR

The proposed web prototype provides a multi-centric system to access to EHR information. Oncoming work is devoted to apply it in the daily practice for automating and streamlining the clinicians workflow, with the long-term ability to generate a complete record of a clinical patient encounter directly from the information of the EHR. Towards that end, the proposed indices in the literature, such as HRV or T-wave alternans, require a much wider and complex knowledge domains specification, and further effort is to be made for their definition. Consequently, they will be addressed as future work. On the other hand, new tools, not only based on logical relationships but also on advance ML and data mining techniques will be used to evidence-based decision support.

Bibliography

- [1] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, “An overview of machine learning,” in *Machine learning*. Springer, 1983, pp. 3–23.
- [2] J. R. Anderson, R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Morgan Kaufmann, 1986, vol. 2.
- [3] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
- [4] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. Springer New York, 2006, vol. 4, no. 4.
- [5] V. Cherkassky and F. M. Mulier, *Learning from data: concepts, theory, and methods*. John Wiley & Sons, 2007.
- [6] I. Guyon, S. Gunn, M. Nikravesh, and L. Z. (Ed.), *Feature extraction: foundations and applications*. Heidelberg: Springer, 2006.
- [7] F. Wang, “Data analytics in healthcare: Problems, challenges and future directions,” in *Information and Knowledge Management*, 2014.
- [8] I. Guyon and A. Elisseeff, “An introduction to feature extraction,” in *Feature Extraction*. Springer, 2006, pp. 1–25.
- [9] S. Mori, H. Nishida, and H. Yamada, *Optical character recognition*. John Wiley & Sons, Inc., 1999.
- [10] L. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

-
- [11] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*. PTR Prentice Hall Englewood Cliffs, 1993, vol. 14.
- [12] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab, 1999.
- [13] M. J. Berry and G. Linoff, *Data mining techniques: for marketing, sales, and customer support*. John Wiley & Sons, Inc., 1997.
- [14] M. J. Shaw, C. Subramaniam, G. W. Tan, and M. E. Welge, “Knowledge management and data mining for marketing,” *Decision support systems*, vol. 31, no. 1, pp. 127–137, 2001.
- [15] R. Blatterg, R. Briescha, and E. Fox, “How promotions work,” *Marketing Science*, vol. 14, pp. 122–132, 1995.
- [16] D. Bell, J. Chiang, and V. Padamanabhan, “The decomposition of promotional response: An empirical generalization,” *Marketing Science*, vol. 18, no. 4, pp. 504–546, 1999.
- [17] P. S. Leeflang and D. R. Wittink, “Building models for marketing decisions:: Past, present and future,” *International journal of research in marketing*, vol. 17, no. 2, pp. 105–126, 2000.
- [18] T. Mitchell, *Machine Learning*. Boston, MA: McGraw-Hill, 1997.
- [19] H. J. V. Heerde, P. S. H. Leeflang, and D. R. Wittink, “Semiparametric analysis to estimate the deal effect curve,” *Journal of Marketing Research*, vol. 38, no. 2, pp. 197–215, May 2001.
- [20] B. Liu, F. Kong, and X. Yang, “Profits estimation in prices promotion,” *In Proc. of 2004 international conf. on Machine Learning and Cybernetics*, vol. 2, pp. 1146–51, 2004.
- [21] M. P. Martínez-Ruiz, A. Mollá-Descals, and J. L. Rojo-Álvarez, “Using daily store-level data to understand price promotion effects in a semiparametric regression model,” *Retailing and Consumer Services*, vol. 3, no. 13, pp. 193–204, 2006.
- [22] M. P. Martínez-Ruiz, A. Mollá-Descals, M. A. Gómez-Borja, and J. L. Rojo-Álvarez, “Evaluating temporary retail price discounts using semiparametric regression,” *Journal of Product & Brand Management*, vol. 15, no. 1, pp. 73–80, 2006.
- [23] M. P. Martínez-Ruiz, A. Mollá-Descals, M. A. Gómez-Borja, and J. L. Rojo-Álvarez, “Evaluating temporary retail price discounts using semiparametric regression,” *Journal of Retailing and Consumer Services*, vol. 3, no. 13, pp. 193–204, 2006.

- [24] M. P. Martínez-Ruiz, A. Mollá-Descals, M. A. Gómez-Borja, and J. L. Rojo-Álvarez, "Using daily store-level data to understand price promotion effects in a semiparametric regression model," *Journal of Retailing and Consumer Services*, vol. 13, no. 3, pp. 193–204, 2006.
- [25] T. Wang, Y. Li, and S. Zhao, "Application of SVM based on rough set in real estate prices prediction," *4th Intl. Conf. on Wireless Communications, Networking and Mobile Computing*, pp. 1–4, 2008.
- [26] J. R. (Ed.), *Health Information Systems: Concepts, Methodologies, Tools, and Applications*. Hershey New York: IGI Global, 2010.
- [27] H. N. Nguyen, S. Y. Ohn, J. Park, and K. S. Park, "Combined kernel function approach in svm for diagnosis of cancer," in *Advances in Natural Computation*. Springer, 2005, pp. 1017–1026.
- [28] J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, and et al., "Heterogeneous data fusion for alzheimer's disease study," in *Proceedings of the 14th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*. ACM, 2008, pp. 1025–1033.
- [29] S. DeLisle, B. South, J. A. Anthony, E. Kalp, A. Gundlapalli, F. C. Curriero, G. E. Glass, M. Samore, and T. M. Perl, "Combining free text and structured electronic medical record entries to detect acute respiratory infections," *PLOS ONE*, vol. 5, no. 10, p. e13377, 2010.
- [30] J. Buckley, S. Coopey, J. Sharko, F. Polubriaginof, B. Drohan, A. Belli, E. H. Kim, J. Garber, B. Smith, M. Gadd, M. Specht, C. Roche, T. Gudewicz, and K. Hughes, "The feasibility of using natural language processing to extract clinical information from breast pathology reports," *Journal of Pathology Informatics*, vol. 3, no. 23, pp. 1–7, 2012.
- [31] Y. Wang, Z. Yua, Y. Jiangb, Y. Liuc, L. Chena, and Y. Liua, "A framework and its empirical study of automatic diagnosis of traditional chinese medicine utilizing raw free-text clinical records," *Journal of biomedical informatics*, vol. 45, no. 2, pp. 210–223, 2012.
- [32] A. Wright, A. McCoy, S. Henkin, A. Kale, and D. Sittig, "Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions," *Journal of the American Medical Informatics Association*, vol. 20, no. 5, pp. 887–890, 2013.
- [33] Z. Zhang, Y. Xu, J. Liu, D. W. K. Wong, C. K. Kwoh, and et al., "Automatic diagnosis of pathological myopia from heterogeneous biomedical data," *PLOS ONE*, vol. 8, no. 6, p. e65736, 2013.

- [34] I. Larsen, *Cancer in Norway 2011 - Cancer incidence, mortality, survival and prevalence in Norway*. Oslo: Cancer Registry of Norway: Cancer Registry of Norway, 2013. [Online]. Available: <http://www.kreftregisteret.no/no/Generelt/Nyheter/Nokkeltall---kreft-2011/>
- [35] H. Kehlet, “Fast-track colorectal surgery,” *The Lancet*, vol. 371, no. 9615, pp. 791–3, 2008.
- [36] H. Snijders, D. Henneman, N. van Leersum, M. T. Berge, M. Fiocco, T. Karsten, K. Havenga, T. Wiggers, J. Dekker, R. Tollenaar, and M. Wouters, “Anastomotic leakage as an outcome measure for quality of colorectal cancer surgery,” *BMJ quality & safety*, vol. 22, no. 9, pp. 759–67, 2013.
- [37] N. Hirst, J. Tiernan, P. Millner, and D. Jayne, “Systematic review of methods to predict and detect anastomotic leakage in colorectal surgery,” *Colorectal Dis*, 2013.
- [38] A. Karliczek, N. Harlaar, C. Zeebregts, T. Wiggers, P. Baas, and G. van Dam, “Surgeons lack predictive accuracy for anastomotic leakage in gastrointestinal surgery,” *International Journal of Colorectal Disease*, vol. 24, no. 5, pp. 569–76, 2009.
- [39] J. Dekker, G. Liefers, J. de Mol van Otterloo, H. Putter, and R. Tollenaar, “Predicting the risk of anastomotic leakage in left-sided colorectal surgery using a colon leakage score,” *Journal of Surgical Research*, vol. 166, no. 1, pp. e27 – e34, 2011.
- [40] D. M. Lloyd-Jones, “Cardiovascular risk prediction: Basic concepts, current status, and future directions,” *Circulation*, vol. 121, pp. 1768–77, 2010.
- [41] S. Kotsiantis, “Supervised machine learning: A review of classification techniques,” *Informatica*, vol. 31, pp. 249–268, 2007.
- [42] S. Theodoridis and K. Koutroubas, *Pattern Recognition*. Academic Press, 2007.
- [43] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, “Missing value estimation methods for dna microarrays,” *Bioinformatics*, vol. 17, no. 6, pp. 520–525, 2001.
- [44] T. A. Lasko, J. C. Denny, and M. A. Levy, “Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data,” *PLOS ONE*, vol. 8, no. 6, p. e66341, 2013.
- [45] B. Efron and R. Tibshirani, *An introduction to the bootstrap*. Boca Raton, FL: Chapman & Hall, 1997.
- [46] R. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, pp. 179–88, 1936.

- [47] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, vol. 129, pp. 103–30, 1997.
- [48] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, pp. 131–63, 1997.
- [49] L. Devroye, L. Györfi, and G. Lugosi, *Probabilistic theory of pattern recognition. Stochastic modeling and applied probability*. New York: Springer, 1996.
- [50] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.
- [51] S. J. Wright and J. Nocedal, *Numerical optimization*. Springer New York, 1999, vol. 2.
- [52] P. D. Wasserman, *Advanced methods in neural computing*. John Wiley & Sons, Inc., 1993.
- [53] V. Vapnik, *Statistical Learning Theory*. John Wiley and Sons, Inc., New York, 1998.
- [54] B. Schölkopf and A. J. Smola, *Learning with kernels*. Cambridge, MA: MIT Press, 2002.
- [55] B. Schölkopf, A. Smola, R. Williamson, and P. Bartlett, "New support vector algorithms," *Neural computation*, vol. 12, no. 5, pp. 1207–45, 2000.
- [56] H. T. Lin, C. J. Lin, and R. C. Weng, "A note on platt's probabilistic outputs for support vector machines," *Machine learning*, vol. 68, no. 3, pp. 267–276, 2007.
- [57] C. C. Chang and C. J. Lin, "Training v-support vector regression: theory and algorithms," *Neural Computation*, vol. 14, no. 8, pp. 1959–1977, 2002.
- [58] G. Camps-Valls and L. Bruzzone, *Kernel methods for remote sensing data analysis*. Wiley Online Library, 2009, vol. 26.
- [59] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *The Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.
- [60] M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy, "Composite kernel learning," *Machine learning*, vol. 79, no. 1-2, pp. 73–103, 2010.
- [61] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.
- [62] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Proceedings of the 14th Annual Conference on Neural Information Processing Systems*, Montreal, Canada, 2001, pp. 682–688.

- [63] M. Girolami, "Orthogonal series density estimation and the kernel eigenvalue problem," *Neural Computation*, vol. 14, no. 3, pp. 669–688, 2002.
- [64] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet, "Learning eigenfunctions links spectral embedding and kernel pca," *Neural Computation*, vol. 16, no. 10, pp. 2197–2219, 2004.
- [65] S. Vijayakumar and S. Schaal, "Locally weighted projection regression: An algorithm for incremental real time learning in high dimensional space," in *The Sixteenth Intl. Conf. on Machine Learning*,, 2000.
- [66] C. E. Rasmussen, *Gaussian processes for machine learning*. Citeseer, 2006.
- [67] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using gaussian process techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 5, pp. 1832–1843, 2012.
- [68] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1, pp. 389–422, 2002.
- [69] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.
- [70] P. Pavlidis, J. Cai, J. Wetson, and W. Grundy, "Gene functional analysis from heterogeneous data," in *Proc. of the 5th IC on Computation Molecular Biology*, Montreal, QC, Canada, 2000, pp. 242–8.
- [71] J. Brank, M. Grobelnik, N. Milic-Frayling, and D. Mladenic, "Feature selection using support vector machines," in *Proc. of the 3rd Int. Conf. on Data Mining Methods and Databases for Engineering, Finance, and Other Fields*, Bologna, Italy, 2002, pp. 84–89.
- [72] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *Proc. of the 14th ICML97*, San Francisco, CA, USA, 1997, pp. 412–320.
- [73] D. Mladenic and M. Grobelnik, "Feature selection for unbalanced class distribution and Naive Bayes," in *Proc. of the 16th ICML99*, Bled, Slovenia, 1999, pp. 258–67.
- [74] G. Forman, "An experimental study of feature selection metrics for text categorization," *The Journal of Machine Learning Research*, vol. 3, pp. 1289–305, 2003.
- [75] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial intelligence*, vol. 151, no. 1, pp. 155–176, 2003.

- [76] G. Hughes, "On the mean accuracy of statistical pattern recognition," *IEEE Trans. Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.
- [77] K. Fukunaga, "Effect of sample size in classifier design," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 8, pp. 873–85, 1989.
- [78] G. Cawley and N. Talbot, "Fast exact leave-one-out cross-validation of sparse least-squares support vector machines," *Neural Networks*, vol. 17, pp. 1467–75, 2004.
- [79] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
- [80] A. Renyi, "On measures of entropy and information," in *Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1960, pp. 547–561.
- [81] R. Jenssen, "Kernel entropy component analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 847–860, 2010.
- [82] H. M. Park, "Comparing group means: t-tests and one-way anova using stata, sas, r, and spss," Technical Working Paper. The University Information Technology Services (UITS) Centre for Statistical and Mathematical Computing, Indiana University, Tech. Rep., 2003.
- [83] J. Peacock and P. Peacock, *Oxford Handbook of Medical Statistics*. Oxford, UK: Oxford University Press Print, 2010.
- [84] S. Geisser and W. F. Eddy, "A predictive approach to model selection," *Journal of the American Statistical Association*, vol. 74, no. 365, pp. 153–160, 1979.
- [85] S. Haykin and R. Lippmann, "Neural networks, a comprehensive foundation," *International Journal of Neural Systems*, vol. 5, no. 4, pp. 363–364, 1994.
- [86] J. L. Rojo-Álvarez, A. Arenal-Maiz, and A. Artes-Rodríguez, "Discriminating between supraventricular and ventricular tachycardias from egm onset analysis," *IEEE Engineering in Medicine and Biology Magazine*, vol. 21, no. 1, pp. 16–26, 2002.
- [87] J. Quelch, "How to market in a recession," *Marketing KnowHow*, September 2008.
- [88] R. Blattberg and A. Neslin, *Sales promotion: concepts, methods and strategies*. Prentice-Hall, 1990.
- [89] K. Philip, "Marketing management: analysis planning implementation and control," 2005.
- [90] T. Yeshin, *Sales promotion*. Cengage Learning EMEA, 2006.

- [91] R. C. Goodstein, "Marketing for the entrepreneur: Customer focus to multiple constituencies," *Entrepreneurship and Economic Growth in the American Economy*, vol. 12, pp. 193–208, 2000.
- [92] A. Krishna, "The impact of dealing patterns on purchase behavior," *Marketing Science*, vol. 13, no. 4, pp. 351–373, 1994.
- [93] V. Kumar and A. Pereira, "Assessing the competitive impact of type, timing, frequency, and magnitude of retail promotions," *Journal of Business Research*, vol. 40, no. 1, pp. 1–13, 1997.
- [94] V. Shankar and R. N. Bolton, "An empirical analysis of determinants of retailer pricing strategy," *Marketing Science*, vol. 23, no. 1, pp. 28–49, 2004.
- [95] S. J. Hoch, X. Dreze, and M. E. Purk, "EDLP, Hi-Lo, and margin arithmetic," *The Journal of Marketing*, pp. 16–27, 1994.
- [96] L. Rajiv and R. Rao, "Supermarket competition: The case of everyday low price," *Marketing Science*, vol. 16, no. 1, pp. 60–80, 1997.
- [97] G. Voss and K. Seiders, "Exploring the effect of retail sector and firm characteristics on retail price promotion strategy," *Journal of Retailing*, vol. 79, no. 1, pp. 37–52, 2003.
- [98] H. J. V. Heerde, P. S. H. Leeflang, and D. R. Wittink, "Semiparametric analysis to estimate the deal effect curve," *Journal of Marketing Research*, vol. 38, no. 2, pp. 197–215, 2001.
- [99] E. A. Blair and E. L. Landon Jr, "The effects of reference prices in retail advertisements," *The Journal of Marketing*, pp. 61–69, 1981.
- [100] H. J. Van Heerde, P. S. Leeflang, and D. R. Wittink, "Decomposing the sales promotion bump with store data," *Marketing Science*, vol. 23, no. 3, pp. 317–334, 2004.
- [101] T. Hastie and R. Tibshirani, *Generalized additive models*. Chapman & Hall/CRC, 1990.
- [102] D. Ruppert, M. Wand, and R. Carroll, *Semiparametric regression*. Cambridge Univ. Press, 2003.
- [103] P. Kopalle, C. Mela, and L. Marsh, "The dynamic effect of discounting on sales: Empirical analysis and normative pricing implications," *Marketing Science*, vol. 18, no. 3, pp. 317–332, 1999.
- [104] T. Ando, "Bayesian state space modeling approach for measuring the effectiveness of marketing activities and baseline sales from POS data," in *In Proc. of IEEE Int. Conf. on Data Mining*, Las Vegas, USA, Jun 2006, pp. 21–32.

- [105] C. Cuthbertson and J. Reynolds, *Retail Strategy*. Oxford: Routledge, 2012.
- [106] R. Rooderkerk, H. Van Heerde, and T. H. A. Bijmolt, “Optimizing retail assortments,” *Marketing Science*, vol. 32, no. 5, pp. 699–715, 2013.
- [107] Ó. González-Benito, M. Martínez-Ruiz, and A. Mollá-Descals, “Spatial mapping of price competition using logit-type market share models and store-level scanner-data,” *Journal of the Operational Research Society*, vol. 60, no. 1, pp. 52–62, 2009.
- [108] Ó. González-Benito, M. P. Martínez-Ruiz, and A. Mollá-Descals, “Using store level scanner data to improve category management decisions: Developing positioning maps,” *European Journal of Operational Research*, vol. 198, no. 2, pp. 666–674, 2009.
- [109] —, “Retail pricing decisions and product category competitive structure,” *Decision Support Systems*, vol. 49, no. 1, pp. 110–119, 2010.
- [110] M. Ataman, H. Van Heerde, and C. Meta, “The long-term effects of marketing strategy on brand sales,” *Journal of Marketing Research*, vol. 47, no. 5, pp. 866–882, 2010.
- [111] I. Alon, M. Qi, and R. Sadowski, “Forecasting aggregate retail sales: A comparison of artificial neural networks and traditional methods,” *Journal of Retailing and Consumer Services*, vol. 8, pp. 147–156, 2001.
- [112] G. Zotterria and M. Kalchschmidt, “A model for selecting the appropriate level of aggregation in forecasting processes,” *International Journal Production Economics*, vol. 108, pp. 74–83, 2007.
- [113] A. Bodapati and S. Gupta, “The recoverability of segmentation structure from store-level aggregate data,” *Journal of Marketing Research*, vol. 41, no. 3, pp. 351–64, 2004.
- [114] E. Foekens, P. Leeflang, and D. Wittink, “A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation,” *International Journal of Forecasting*, vol. 10, no. 2, pp. 245–61, 1994.
- [115] M. Christen, S. Gupta, J. C. Porter, R. Staelin, and D. R. Wittink, “Using market-level data to understand promotion effects in a nonlinear mode,” *Journal of Marketing Research*, vol. 34, no. 3, pp. 322–334, 1997.
- [116] T. Steven, “Avoiding aggregation bias in demand estimation: A multivariate promotional disaggregation approach,” *Quantitative Marketing and Economics*, vol. 4, pp. 383–405, 2006.

- [117] R. Link, "Are aggregate scanner data models biased?" *Journal of Advertising Research*, 1995.
- [118] G. Marnik and M. Dominique, "Time-series models in marketing: Past, present and future," *International Journal of Research in Marketing*, vol. 17, pp. 183–193, 2000.
- [119] C. Horvath, M. Kornelis, and P. S. H. Leeflang, "What marketing scholars should know about time series analysis : Time series applications in marketing," *University of Groningen, Research Institute, Systems, Organizations and Management, Research Report*, 2002.
- [120] P. Brockwell and R. Davis, *Introduction to time series and forecasting*. New York: Springer, 2002.
- [121] K. Jonassonm and S. E. Ferrando, "Evaluating exact varma likelihood and its gradient when data are incomplete," *ACM Transactions on Mathematical Software*, vol. 35, no. 1, 2008.
- [122] S. Garde, P. Knaup, E. Hovenga, and S. Heard, "Towards semantic interoperability for electronic health records: domain knowledge governance for openEHR archetypes," *Methods of Information in Medicine*, vol. 46, no. 3, pp. 332–43, 2007.
- [123] S. S. Magill, W. Hellinger, J. Cohen, R. Kay, C. Bailey, B. Boland, D. Carey, J. d. Guzman, K. Dominguez, J. Edwards *et al.*, "Prevalence of healthcare-associated infections in acute care hospitals in Jacksonville, Florida," *Infection Control*, vol. 33, no. 03, pp. 283–291, 2012.
- [124] G. de Lissovoy, K. Fraeman, V. Hutchins, D. Murphy, D. Song, and B. B. Vaughn, "Surgical site infection: incidence and impact on hospital utilization and treatment costs," *American Journal of Infection Control*, vol. 37, no. 5, pp. 387–397, 2009.
- [125] P. L. Owens, M. L. Barrett, S. Raetzman, M. Maggard-Gibbons, and C. A. Steiner, "Surgical site infections following ambulatory surgery procedures," *JAMA*, vol. 311, no. 7, pp. 709–716, 2014.
- [126] E. H. Lawson, B. L. Hall, and C. Y. Ko, "Risk factors for superficial vs deep/organ-space surgical site infections: implications for quality improvement initiatives," *JAMA surgery*, vol. 148, no. 9, pp. 849–858, 2013.
- [127] E. H. Lawson, C. Y. Ko, J. L. Adams, W. B. Chow, and B. L. Hall, "Reliability of evaluating hospital quality by colorectal surgical site infection type," *Annals of surgery*, vol. 258, no. 6, pp. 994–1000, 2013.

- [128] J. Blumetti, M. Luu, G. Sarosi, K. Hartless, J. McFarlin, B. Parker, S. Dineen, S. Huerta, M. Asolati, E. Varela *et al.*, “Surgical site infections after colorectal surgery: do risk factors vary depending on the type of infection considered?” *Surgery*, vol. 142, no. 5, pp. 704–711, 2007.
- [129] Z. Afzal, M. Engelkes, K. Verhamme, H. Janssens, M. Sturkenboom, J. Kors, and M. Schuemie, “Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases,” *Pharmacoepidemiol Drug Saf*, vol. 22, no. 8, pp. 826–33, 2013.
- [130] Z. Wang, A. Shah, A. Tate, S. Denaxas, J. Shawe-Taylor, and H. Hemingway, “Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning,” *PLOS ONE*, vol. 7, no. 1, pp. 1–9, 2012.
- [131] R. Cobb, S. Puri, D. Z. Wang, T. Baslanti, and A. Bihorac, “Knowledge extraction and outcome prediction using medical notes,” in *ICML workshop on Role of Machine Learning in Transforming Healthcare*, Atlanta, Georgia, USA, 2013.
- [132] P. Nadkarni, L. Ohno-Machado, and W. Chapman, “Natural language processing: an introduction,” *Journal of the American Medical Informatics Association*, vol. 18, pp. 544–51, 2011.
- [133] S. Meystre, G. Savova, K. Kipper-Schule, and J. Hurdle, “Extracting information from textual documents in the electronic health record: a review of recent research,” *Yearbook of Medical Informatics*, pp. 128–44, 2008.
- [134] M. Khalilia, S. Chakraborty, and M. Popescu, “Predicting disease risks from highly imbalanced data using random forest,” *BMC Medical Informatics and Decision Making*, vol. 11, no. 51, pp. 1–13, 2011.
- [135] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in *Machine Learning: ECML-98, Lecture Notes in Computer Science*, vol. 1398, Chemnitz, Germany, 1998, pp. 137–42.
- [136] H. Drucker, D. Wu, and V. Vapnik, “Support vector machines for spam categorization,” *IEEE Trans Neural Networks*, vol. 10, no. 5, pp. 1048–54, 1999.
- [137] E. Leopold and J. Kindermann, “Text categorization with support vector machines. How to represent texts in input space?” *Machine Learning*, vol. 46, pp. 423–44, 2002.

- [138] A. Graf, A. Smola, and S. Borer, "Classification in a normalized feature space using support vector machines," *IEEE Trans Neural Networks and Learning Systems*, vol. 14, no. 3, pp. 597–605, 2003.
- [139] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–23, 1988.
- [140] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–84, 2009.
- [141] B. Silverman, *Density Estimation for Statistics and Data Analysis*. London, UK: Chapman & Hall CRC, 1986.
- [142] C. Shivade, P. Raghava, E. Fosler-Lussier, P. Embi, N. Elhadad, S. Johnson, and A. Lai, "A review of approaches to identifying patient phenotype cohorts using electronic health records," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 221–30, 2014.
- [143] C. Paxton, A. Niculescu-Mizil, and S. Saria, "Developing predictive models using electronic medical records: challenges and pitfalls," in *AMIA Annual Symposium Proceedings*, vol. 2013. American Medical Informatics Association, 2013, p. 1109.
- [144] C. Soguero-Ruiz, K. Hindberg, J. L. Rojo-Álvarez, S. O. Skrovseth, F. Godtlielsen, K. Mortensen, A. Revhaug, R.-O. Lindsetmo, I. Mora-Jiménez, K. M. Augestad *et al.*, "Bootstrap resampling feature selection and support vector machine for early detection of anastomosis leakage," in *Intl. Conf. IEEE EMBS Biomedical and Health Informatics*, 2014, pp. 577–580.
- [145] C. Soguero-Ruiz, K. Hindberg, J. Rojo-Álvarez, S. Skrovseth, F. Godtlielsen, K. Mortensen, A. Revhaug, R.-O. Lindsetmo, K. Augestad, and R. Jenssen, "Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, Accepted for publication October 2014.
- [146] N. E. El Faouzi, H. Leung, and A. Kurian, "Data fusion in intelligent transportation systems: Progress and challenges - a survey," *Information Fusion*, vol. 12, no. 1, pp. 4–10, 2011.
- [147] S. G. Iyengar, P. K. Varshney, and T. Damarla, "A parametric copula-based framework for hypothesis testing using heterogeneous data," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2308–2319, 2011.

- [148] G. Camps-Valls, L. Gomez-Chova, J. Muñoz-Marí, J. Vila-Francés, and J. Calpe-Maravilla, “Composite kernels for hyperspectral image classification,” *Geoscience and Remote Sensing Letters, IEEE*, vol. 3, no. 1, pp. 93–97, 2006.
- [149] T. Joachims, N. Cristianini, and J. Shawe-Taylor, “Composite kernels for hypertext categorisation,” in *Proc. of the Eighteenth Intl. Conf. on Machine Learning*, vol. 1, San Francisco, CA, USA, 2001, pp. 250–57.
- [150] D. Tuia, F. Ratle, A. Pozdnoukhov, and G. Camps-Valls, “Multisource composite kernels for urban-image classification,” *Geoscience and Remote Sensing Letters, IEEE*, vol. 7, no. 1, pp. 88–92, 2010.
- [151] M. T. Bahadori and Y. Liu, “Granger causality analysis in irregular time series.” in *Secure Data Management*, 2012, pp. 660–671.
- [152] L. Erb, N. H. Hyman, and T. Osler, “Abnormal vital signs are common after bowel resection and do not predict anastomotic leak,” *Journal of the American College of Surgeons*, vol. 218, no. 6, pp. 1195–1199, 2014.
- [153] Z. Wu, D. Freek, and J. Lange, “Do normal clinical signs and laboratory tests exclude anastomotic leakage?” *Journal of the American College of Surgeons*, vol. 219, no. 1, p. 164, 2014.
- [154] M. den Dulk, S. L. Noter, E. R. Hendriks, M. A. Brouwers, C. H. van der Vlies, and et al., “Improved diagnosis and treatment of anastomotic leakage after colorectal surgery,” *European Journal of Surgical Oncology*, vol. 35, no. 4, pp. 420–426, 2009.
- [155] S. O. Skrøvseth, K. M. Augestad, and S. Ebadollahi, “Data-driven approach for assessing utility of medical tests using electronic medical records,” *Journal of Biomedical Informatics*, vol. 53, no. 0, pp. 270 – 276, 2015.
- [156] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.
- [157] P. Singh, I. Zeng, S. Srinivasa, D. Lemanu, A. Connolly, and A. Hill, “Systematic review and meta-analysis of use of serum c-reactive protein levels to predict anastomotic leak after colorectal surgery,” *British Journal of Surgery*, vol. 101, no. 4, pp. 339–346, 2014.
- [158] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, “Model-building strategies and methods for logistic regression,” *Applied Logistic Regression, Third Edition*, pp. 89–151, 2000.

- [159] C. Paxton, A. Niculescu-Mizil, and S. Saria, “Developing predictive models using electronic medical records: challenges and pitfalls,” in *AMIA Annual Symposium Proceedings*, vol. 2013. Washington DC: American Medical Informatics Association, 2013, p. 1109.
- [160] D. H. Culver, T. C. Horan, R. P. Gaynes, W. J. Martone, W. R. Jarvis, T. G. Emori, S. N. Banerjee, J. R. Edwards, J. S. Tolson, T. S. Henderson *et al.*, “Surgical wound infection rates by wound class, operative procedure, and patient risk index,” *The American Journal of Medicine*, vol. 91, no. 3, pp. S152–S157, 1991.
- [161] C. van Walraven and R. Musselman, “The surgical site infection risk score (ssirs): a model to predict the risk of surgical site infections,” *PLOS ONE*, vol. 8, no. 6, p. e67167, 2013.
- [162] A. Gbegnon, W. N. Street, J. Monestina, and J. W. Cromwell, “Predicting surgical site infections in real-time,” http://cci.drexel.edu/HI-KDD2014/morning_6.pdf, 2010.
- [163] A. A. Gawande, M. R. Kwaan, S. E. Regenbogen, S. A. Lipsitz, and M. J. Zinner, “An Apgar score for surgery,” *Journal of the American College of Surgeons*, vol. 204, no. 2, pp. 201–208, 2007.
- [164] V. A. Constantinides, P. P. Tekkis, A. Senapati, A. of Coloproctology of Great Britain *et al.*, “Comparison of POSSUM scoring systems and the surgical risk scale in patients undergoing surgery for complicated diverticular disease,” *Diseases of the colon & rectum*, vol. 49, no. 9, pp. 1322–1331, 2006.
- [165] K. G. Cologne, D. S. Keller, L. Liwanag, B. Devaraj, and A. J. Senagore, “Use of the American College of Surgeons NSQIP surgical risk calculator for laparoscopic colectomy: How good is it and how can we improve it?” *Journal of the American College of Surgeons*, vol. 220, no. 3, 2015.
- [166] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Campbell, and D. W. Bates, “Grand challenges in clinical decision support,” *Journal of Biomedical Informatics*, vol. 41, no. 2, pp. 387–92, 2008.
- [167] A. Wright, D. W. Bates, B. Middleton, T. Hongsermeier, V. Kashyap, S. M. Thomas, and D. F. Sittig, “Creating and sharing clinical decision support content with Web 2.0: Issues and examples,” *Journal of Biomedical Informatics*, vol. 42, no. 2, pp. 334–46, 2009.
- [168] M. Marcos, J. A. Maldonado, B. Martínez-Salvador, D. Boscá, and M. Robles, “Interoperability of clinical decision-support systems and electronic health records using archetypes: A case study in clinical trial eligibility,” *Journal of Biomedical Informatics*, vol. 46, no. 4, pp. 676–89, 2013.

- [169] F. Barbarito, F. Pincioli, J. Mason, S. Marceglia, L. Mazzola, and S. Bonacina, "Implementing standards for the interoperability among healthcare providers in the public regionalized healthcare information system of the Lombardy region," *Journal of Biomedical Informatics*, vol. 45, no. 4, pp. 736–45, 2012.
- [170] K. Häyrynen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," *International Journal of Medical Informatics*, vol. 77, no. 5, pp. 291–304, 2008.
- [171] B. E. Dixon, D. J. Vreeman, and S. J. Grannis, "The long road to semantic interoperability in support of public health: Experiences from two states," *Journal of Biomedical Informatics*, vol. 49, pp. 3–8, 2014.
- [172] O. Bodenreider and A. Burgun, "Biomedical Ontologies," in *Medical Informatics*, 2005, vol. 8, pp. 211–36.
- [173] A. Bauer, M. Malik, G. Schmidt, P. Barthel, H. Bonnemeier, and et al., "Heart rate turbulence: standards of measurement, physiological interpretation, and clinical use: (ishne consensus)," *Journal of the American College of Cardiology*, vol. 52, no. 17, pp. 1353–65, 2008.
- [174] (2015, Jun) Unified Medical Language System (UMLS). Available from: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html.
- [175] G. Héja, S. György, and P. Varga, "Ontological Analysis of SNOMED CT," *BMC Medical Informatics and Decision Making*, 8 (Suppl 1): S8, 2008.
- [176] C. Wroe, "Is Semantic Web Technology Ready for Healthcare?" in *Proc 3rd Annual European Semantic Web Conference*, A. Leger, A. Kulas, L. Nixon, and R. Meersman, Eds., Budva, Montenegro, Jun 2006.
- [177] D. Truran, P. Saad, M. Zhang, K. Innes, M. Kemp, S. Huckson, and S. Bennetts, "Using SNOMED CT(R) - Enabled Data Collections in a National Clinical Research Program: Primary Care Data Can Be Used in Secondary Studies," in *Proc Frontiers of Health Informatics - Redefining Healthcare*, V. Sintchenko and P. Croll, Eds., Aug 2009.
- [178] "SNOMED-CT. International Health Terminology Standards Development Organisation," Available from: <http://www.ihtsdo.org/snomed-ct/>.
- [179] "EN 13606 association," Available from: <http://www.en13606.org>.
- [180] "CEN/TC251. EN13606-1- Medical informatics-electronic health record communication. Part 1. Reference Model."

- [181] “CEN/TC251, EN13606-2-Health informatics-electronic health record communication. Part 2. Archetypes.”
- [182] J. L. Rojo-Álvarez, O. Barquero-Pérez, I. Mora-Jiménez, E. Everss, A. B. Rodriguez-Gonzalez, and A. Garcia-Alberola, “Heart rate turbulence denoising using support vector machines,” *IEEE Trans. Biomedical Engineering*, vol. 56, no. 2, pp. 310–19, 2009.
- [183] M. A. Watanabe, “Heart Rate Turbulence: A Review,” *J. Indian Pacing Electrophysiol*, vol. 3, no. 1, pp. 10–22, 2003.
- [184] D. Smith, K. Solem, P. Laguna, J. Martínez, and L. Sörnmo, “Model-Based Detection of Heart Rate Turbulence Using Mean Shape Information,” *IEEE Trans. Biomedical Engineering*, vol. 57, no. 2, pp. 334–42, 2010.
- [185] J. Martínez, I. Cygankiewicz, D. Smith, A. Bayes de Luna, P. Laguna, and L. Sörnmo, “Detection Performance and Risk Stratification Using a Model-Based Shape Index Characterizing Heart Rate Turbulence,” *Annals of Biomedical Engineering*, vol. 38, no. 10, pp. 3173–84, 2010.
- [186] M. A. Watanabe and G. Schmidt, “Heart Rate Turbulence: A 5-year review,” *Heart Rhythm*, vol. 1, no. 6, pp. 732–8, 2004.
- [187] (2015) Foundational Model of Anatomy. University of Washington. Available from: <http://sig.biostr.washington.edu/projects/fm/AboutFM.html>.
- [188] A. Jovic, M. Prcela, and D. Gamberger, “Ontologies in Medical Knowledge Representation,” in *Proc 29th Int. Conf. on Information Technology Interfaces*, Cavtat, Croatia, June 2007, pp. 535–40.
- [189] B. Gonçalves, G. Guizzardi, and J. G. Pereira-Filho, “An Electrocardiogram (ECG) Domain Ontology,” in *2nd Workshop on Ontologies and Metamodels in Software and Data Engineering*, Brazil, Oct 2007, pp. 68–81.
- [190] B. Gonçalves, V. Zamborlini, G. Guizzardi, and J. G. Pereira-Filho, “Using a Lightweight Ontology of Heart Electrophysiology in an Interactive Web Application,” in *Proc. XIV Brazilian Symposium on Multimedia and the Web*, Brazil, 2008, pp. 77–80.
- [191] G. Acampora, C. S. Lee, A. Vitiello, and M. H. Wang, “Evaluating cardiac health through semantic soft computing techniques,” *Soft Computing*, vol. 16, no. 7, pp. 1165–81, 2012.

- [192] T. Tanantong, E. Nantajeewarawat, and S. Thiemjarus, "Towards Continuous Electrocardiogram Monitoring Based on Rules and Ontologies," in *Proc 11th IEEE Intl Conf on Bioinformatics and Bioengineering*, Taichung, Taiwan, Oct 2011, pp. 327–30.
- [193] T. Sampalli, M. Shepherd, and J. Duffy, "A patient profile ontology in the heterogeneous domain of complex and chronic health conditions," in *Proc 44th Hawaii Intl Conf on System Sciences*, Hawaii, Jan 2011, pp. 1–10.
- [194] J. Julina and D. Thenmozhi, "Ontology Based EMR for Decision Making in Health Care using SNOMED CT," in *Proc Intl Conf on Recent Trends in Information Technology*, Chennai, India, Apr 2012, pp. 514–19.
- [195] (2015) CliniClue Xplore. Developed by The Clinical Information Consultancy Ltd. Available from: <http://www.cliniclue.com/>.
- [196] "Archetype Definition Language ADL 1.5," Available from: http://www.openehr.org/downloads/ADLworkbench/learning_about.
- [197] M. Meizoso García, J. L. I. Allones, D. Martínez Hernández, and M. J. Taboada Iglesias, "Semantic similarity-based alignment between clinical archetypes and SNOMED CT: An application to observations," *International Journal of Medical Informatics*, vol. 81, pp. 566–78, 2012.
- [198] O. Lassila and R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," Available from: <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222>, W3C Recommendation 22 Feb. 1999.
- [199] T. Bray, J. Paoli, C. Sperberg-McQueen, E. Maler, and F. Yergeau. (W3C Recommendation 26 Nov. 2008) Extensible Markup Language (XML) 1.0 (Fifth Edition). Available from: <http://www.w3.org/TR/xml/>.
- [200] (2015) World Wide Web Consortium (W3C). Available from: <http://www.w3.org>.
- [201] D. McGuinness and F. Van Harmelen, "OWL Web Ontology Language Overview," Available from: <http://www.w3.org/TR/owl-features>, W3C Recommendation 10 Feb. 2004.
- [202] N. F. Noy and D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Systems Laboratory and Stanford Medical Informatics, Tech. Rep. KSL-01-05 and SMI-2001-0880, 2001.
- [203] Available from: <http://www.sanidadmadrid.org:8989/WebOntoHRT/jsp/index.html>.

- [204] I. Horrocks, P. Patel-Schneider, H. Boley, S. Tabet, B. Grosz, and M. Dean, “SWRL: A Semantic Web Rule Language Combining OWL and RuleML,” National Research Council of Canada, Network Inference, and Stanford University. Available from: <http://www.w3.org/Submission/SWRL/>, W3C Member Submission 21 May 2004.
- [205] J.W. Lloyd, *Foundations of Logic Programming*. Springer-Verlag, 1987.
- [206] J. Alzueta and J. Fernández, “Spanish implantable cardioverter-defibrillator registry. Eighth official report of the Spanish society of cardiology working group on implantable cardioverter-defibrillator,” *Revista Española de Cardiología*, vol. 65, no. 11, pp. 1019–29, 2012.
- [207] J. Koyama, J. Watanabe, A. Yamada, and et al., “Evaluation of heart-rate turbulence as a new prognostic marker in patients with chronic heart failure.” *Circulation Journal*, vol. 66, no. 10, pp. 902–7, 2002.
- [208] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosz, and M. Dean, “SWRL: A Semantic Web Rule Language Combining OWL and RuleML,” p. 79, 2004.
- [209] C. Soguero-Ruiz, L. Lechuga-Suárez, I. Mora-Jiménez, J. Ramos-López, O. Barquero-Pérez, A. García-Alberola, and J. L. Rojo-Álvarez, “Ontology for heart rate turbulence domain from the conceptual model of SNOMED-CT.” *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 7, pp. 1825–33, 2013.
- [210] Commission recommendation of 2 July 2008 on cross-border interoperability of electronic health record systems (notified under document number (C(2008) 3282). Official Journal of the European Union, 2008.
- [211] “OpenEHR consortium,” Available from: <http://www.openehr.org>.
- [212] E. Sundvall, R. Qamar, M. Nystrom, and et. al, “Integration of tools for binding archetypes to SNOMED CT,” *BMC Medical Informatics and Decision Making*, vol. 8, no. Suppl 1, p. S7, 2008.
- [213] S. Yu, D. Berry, and J. Bisbal, “Performance analysis and assessment of a tf-idf based archetype-SNOMED-CT binding algorithm,” in *24th Intl Symp on Computer-Based Medical Systems*, New York, USA, 2011, pp. 1–6.
- [214] “Knowledge management. Template designer,” Available from: https://oceaninformatics.com/solutions/knowledge_management.
- [215] M. Malik, “Heart rate variability: standards of measurement, physiological interpretation, and clinical use,” *Circulation*, vol. 93, no. 5, pp. 1043–65, 1996.

-
- [216] A. Bauer, J. Kantelhardt, P. Barthel, R. Schneider, and et al., “Deceleration Capacity of Heart Rate as a Predictor of Mortality After Myocardial Infarction: Cohort Study,” *Lancet*, vol. 367, no. 9523, pp. 1674–81, 2006.
- [217] J. M. Smith, E. A. Clancy, C. R. Valeri, J. N. Ruskin, and R. J. Cohen, “Electrical Alternans and Cardiac Electrical Instability,” *Circulation*, vol. 77, no. 1, pp. 110–21, 1988.