Healthcare data heterogeneity and its contribution to machine learning performance

PhD DISSERTATION

Advisors
Dr. José Alberto Conejero Casares
Dr. Juan Miguel García Gómez

Author
Francisco Javier Pérez Benito

Valencia, September 2020

A mis padres y a mis hermanos,
por quererme y enseñarme a ser.
A mi mujer,
por aprenderme y, aún así, elegirme.

Acknowledgements

Me tienes tan fascinado.

Fascinado, Sidonie.

Desde mi punto de vista, la vida son etapas y cada etapa se compone de objetivos que cumplir, intereses que conseguir y, por supuesto, personas que te empujan a ello. En un día como hoy, se cierra un ciclo que, afortunadamente, ha iniciado e impulsado otros proyectos en mi vida, con sus nuevos objetivos e intereses. En este sentido, puedo afirmar y afirmo, que doy por cumplidos mis objetivos, por conseguidos mis intereses, y no solo afianzado sino también, ganado la amistad de personas que deseo, me sigan acompañando a lo largo de los futuros retos que me depare la vida. Las siguientes palabras son para esas personas, para las que considero han sido de especial relevancia en este momento de mi vida y a las cuales me esforzaré por mantener siempre cerca de mí.

Sin duda, mi principal hallazgo personal en el desarrollo de esta tesis es haber conocido a mis directores, Alberto y Juanmi. Su ciencia me deslumbró, su carisma me conquistó y su personalidad me atrapó. Durante esta etapa, ha habido momentos dulces, salados y amargos, y en todos ellos me he sentido siempre apoyado. Muchas gracias, sois un referente para mí.

Del mismo modo, quisiera agradecer a mis compañeros del *Biomedical Data Science Lab (BDSLab)* su aceptación y acogida, me sentí uno más desde el primer día y no está de más deciros que extraño el ir al laboratorio y pasar más tiempo con vosotros. En especial, me gustaría mencionar a Carlos Sáez, créeme si te digo que se te aprecia más a tí que al método.

También me gustaría mencionar al *Instituto Tecnológico de la Informática*, en especial al grupo de investigación *Percepción, Reconocimiento, Aprendizaje e Inteligencia Artificial*, no solo por darme la oportunidad de seguir creciendo en el mundo de la ciencia, sino también, por apoyarme en la consecución de mis objetivos personales. François Signol, Juan-Carlos Perez-

Cortes, Rafael Llobet y Javier Cano pienso que también tenéis un papel importante en que yo esté hoy aquí.

Las anteriores personas han aparecido y me han acompañado en el camino que ha supuesto mi doctorado, pero por suerte también hay personas que me han conducido a lo largo de la vida. En primer lugar, me gustaría mencionar a mis amigos, los amigos son esas personas que pese a tu evolución personal y la suya, después de los años están para tí. No importa dónde estemos y cómo de diferentes seamos, sabemos que siempre estamos ahí para el otro. Yo tengo suerte Dani, Bule, Emilio, Víctor, Araceli y Mario. Os tengo. Me tenéis.

Gracias también a Elena, Manolo, Carmen María, Álvaro, Jimena y Mateo. Aunque nuestra relación quizá sea la más reciente de aquellas personas que considero imprescindibles en mi vida, siempre me habéis tranquilizado, apoyado, mimado y dado ánimos. Me siento orgulloso de que forméis parte de mi familia, sé que os sentís orgullosos de mí y espero estar siempre a la altura.

Gracias abuela Ignacia, eres adorable hasta sin quererlo, la cara que me pones cada vez que me ves (aunque lleve barba) me da calor por dentro. Hay personas que te faltan, es un hecho, pero es decisión tuya que cada momento que te han dado te alumbre a lo largo de los días, eso me pasa a mí, me faltan personas, pero no me faltan. Abuelo Jacinto, abuelo Isidoro, abuela Chelo y tito Javi, todos los días me acuerdo de vosotros, y hoy, más. Sé que un día como hoy lo celebraríais incluso más que yo. Os extraño mucho.

Si yo tuviese que elegir mi familia, tengo la suerte de poder decir que sería la misma que me ha tocado. Gracias a mis dos hermanos, Cristina y Óscar, pese a ser más pequeños muchas veces os habéis convertido en faro. Sois buenos, inteligentes e inmejorables. Gracias a mi padre, José Manuel, siempre has sido un ejemplo a seguir, idolatro tu fuerza de voluntad, tu determinación y el apoyo que siempre has sido, gran parte de la culpa de que hoy esté aquí es tuya y de tu cómplice, Puri, mamá. Gracias mamá por tu ternura, inteligencia, bondad y paciencia, intento ser un reflejo tuyo en muchos aspectos de mi vida. Os quiero.

Por último, Carolina, nunca pensé que pudiese, sin necesitarlo, querer estar tanto tiempo con alguien. No simplemente has sido un apoyo continuo en el desarrollo de esta etapa de mi vida, sino que la has complementado y la has hecho más llevadera. Empecé esta tesis siendo tu novio y hoy, a pesar de la tesis, eres mi esposa. Gracias por aguantarme, mimarme, quererme cada día desde que nos conocimos y poner en marcha nuestro proyecto más importante. No me imagino nada sin tí y eres todo lo que necesito para seguir andando.

Table of Contents

Abstract

Somewhere, something incredible is waiting
to be known.

Carl Sagan.

The data quality assessment has many dimensions, from those so obvious as the data completeness and consistency to other less evident such as the correctness or the ability to represent the target population. In general, it is possible to classify them as those produced by an external effect, and those that are inherent in the data itself. This work will be focused on those inherent to data, such as the temporal and the multisource variability applied to healthcare data repositories. Every process is usually improved over time, and that has a direct impact on the data distribution. Similarly, how a process is executed in different sources may vary due to many factors, such as the diverse interpretation of standard protocols by human beings or different previous experiences of experts.

Artificial Intelligence has become one of the most widely extended technological paradigms in almost all the scientific and industrial fields. Advances not only in models but also in *hardware* have led to their use in almost all areas of science. Although the solved problems using this technology often have the drawback of not being interpretable, or at least not as much as other classical mathematical or statistical techniques. This motivated the emergence of the "explainable artificial intelligence" concept, that study methods to quantify and visualize the training process of models based on machine learning.

On the other hand, real systems may often be represented by large networks (graphs), and one of the most relevant features in such networks is the community or clustering structure. Since sociology, biology or clinical situations could usually be modeled using graphs, community detection algorithms are becoming more and more extended in a biomedical field.

In the present doctoral thesis, contributions have been made in the three above mentioned areas. On the one hand, temporal and multisource variabil-

ity assessment methods based on information geometry were used to detect variability in data distribution that may hinder data reuse and, hence, the conclusions which can be extracted from them. This methodology's usability was proved by a temporal variability analysis to detect data anomalies in the electronic health records of a hospital over 7 years.

Besides, it showed that this methodology could have a positive impact if it applied previously to any study. To this end, firstly, we extracted the variables that highest influenced the intensity of headache in migraine patients using machine learning techniques. One of the principal characteristics of machine learning algorithms is its capability of fitting the training set. In those datasets with a small number of observations, the model can be biased by the training sample. The observed variability, after the application of the mentioned methodology and considering as sources the registries of migraine patients with different headache intensity, served as evidence for the truthfulness of the extracted features. Secondly, such an approach was applied to measure the variability among the gray-level histograms of digital mammographies. We demonstrated that the acquisition device produced the observed variability, and after defining an image preprocessing step, the performance of a deep learning model, which modeled a marker of breast cancer risk estimation, increased.

Given a dataset containing the answers to a survey formed by psychometric scales, or in other words, questionnaires to measure psychologic factors, such as depression, cope, etcetera, two deep learning architectures that used the data structure were defined. Firstly, we designed a deep learning architecture using the conceptual structure of such psychometric scales. This architecture was trained to model the happiness degree of the participants, improved the performance compared to classical statistical approaches. A second architecture, automatically designed using community detection in graphs, was not only a contribution to automation but obtained results comparable to its predecessor.

Resumen

El análisis de la calidad de los datos abarca muchas dimensiones, desde aquellas tan obvias como la completitud y la coherencia, hasta otras menos evidentes como la correctitud o la capacidad de representar a la población objetivo. En general, es posible clasificar estas dimensiones como las producidas por un efecto externo y las que son inherentes a los propios datos. Este trabajo se centrará en la evaluación de aquellas inherentes a los datos en repositorios de datos sanitarios, como son la variabilidad temporal y multi-fuente. Los procesos suelen evolucionar con el tiempo, y esto tiene un impacto directo en la distribución de los datos. Análogamente, la subjetividad humana puede influir en la forma en la que un mismo proceso, se ejecuta en diferentes fuentes de datos, influyendo en su cuantificación o recogida.

La inteligencia artificial se ha convertido en uno de los paradigmas tecnológicos más extendidos en casi todos los campos científicos e industriales. Los avances, no sólo en los modelos sino también en el hardware, han llevado a su uso en casi todas las áreas de la ciencia. Es cierto que, los problemas resueltos mediante esta tecnología, suelen tener el inconveniente de no ser interpretables, o al menos, no tanto como otras técnicas de matemáticas o de estadística clásica. Esta falta de interpretabilidad, motivó la aparición del concepto de "inteligencia artificial explicable", que estudia métodos para cuantificar y visualizar el proceso de entrenamiento de modelos basados en aprendizaje automático.

Por otra parte, los sistemas reales pueden representarse a menudo mediante grandes redes (grafos), y una de las características más relevantes de esas redes, es la estructura de comunidades. Dado que la sociología, la biología o las situaciones clínicas, usualmente pueden modelarse mediante grafos, los algoritmos de detección de comunidades se están extendiendo cada vez más en el ámbito biomédico.

En la presente tesis doctoral, se han hecho contribuciones en los tres campos anteriormente mencionados. Por una parte, se han utilizado métodos de evaluación de variabilidad temporal y multi-fuente, basados en geometría de la información, para detectar la variabilidad en la distribución de los datos que pueda dificultar la reutilización de los mismos y, por tanto, las conclusiones que se puedan extraer. Esta metodología demostró ser útil tras ser aplicada a los registros electrónicos sanitarios de un hospital a lo largo de 7 años, donde se detectaron varias anomalías.

Además, se demostró el impacto positivo que este análisis podría añadir a cualquier estudio. Para ello, en primer lugar, se utilizaron técnicas de aprendizaje automático para extraer las características más relevantes, a la hora de clasificar la intensidad del dolor de cabeza en pacientes con migraña. Una

de las propiedades de los algoritmos de aprendizaje automático es su capacidad de adaptación a los datos de entrenamiento, en bases de datos en los que el número de observaciones es pequeño, el estimador puede estar sesgado por la muestra de entrenamiento. La variabilidad observada, tras la utilización de la metodología y considerando como fuentes, los registros de los pacientes con diferente intensidad del dolor, sirvió como evidencia de la veracidad de las características extraídas. En segundo lugar, se aplicó para medir la variabilidad entre los histogramas de los niveles de gris de mamografías digitales. Se demostró que esta variabilidad estaba producida por el dispositivo de adquisición, y tras la definición de un preproceso de imagen, se mejoró el rendimiento de un modelo de aprendizaje profundo, capaz de estimar un marcador de imagen del riesgo de desarrollar cáncer de mama.

Dada una base de datos que recogía las respuestas de una encuesta formada por escalas psicométricas, o lo que es lo mismo cuestionarios que sirven para medir un factor psicológico, tales como depresión, resiliencia, etc., se definieron nuevas arquitecturas de aprendizaje profundo utilizando la estructura de los datos. En primer lugar, se diseño una arquitectura, utilizando la estructura conceptual de las citadas escalas psicométricas. Dicha arquitectura, que trataba de modelar el grado de felicidad de los participantes, tras ser entrenada, mejoró la precisión en comparación con otros modelos basados en estadística clásica. Una segunda aproximación, en la que la arquitectura se diseñó de manera automática empleando detección de comunidades en grafos, no solo fue una contribución de por sí por la automatización del proceso, sino que, además, obtuvo resultados comparables a su predecesora.

Resum

L'anàlisi de la qualitat de les dades comprén moltes dimensions, des d'aquelles tan òbvies com la completesa i la coherència, fins a altres menys evidents com la correctitud o la capacitat de representar a la població objectiu. En general, és possible classificar estes dimensions com les produïdes per un efecte extern i les que són inherents a les pròpies dades. Este treball se centrarà en l'avaluació d'aquelles inherents a les dades en reposadors de dades sanitaris, com són la variabilitat temporal i multi-font. Els processos solen evolucionar amb el temps i açò té un impacte directe en la distribució de les dades. Anàlogament, la subjectivitat humana pot influir en la forma en què un mateix procés, s'executa en diferents fonts de dades, influint en la seua quantificació o arreplega.

La intel·ligència artificial s'ha convertit en un dels paradigmes tecnològics més estesos en quasi tots els camps científics i industrials. Els avanços, no sols en els models sinó també en el maquinari, han portat al seu ús en quasi totes les àrees de la ciència. És cert que els problemes resolts per mitjà d'esta tecnologia, solen tindre l'inconvenient de no ser interpretables, o almenys, no tant com altres tècniques de matemàtiques o d'estadística clàssica. Esta falta d'interpretabilitat, va motivar l'aparició del concepte de "inteligencia artificial explicable", que estudia mètodes per a quantificar i visualitzar el procés d'entrenament de models basats en aprenentatge automàtic.

D'altra banda, els sistemes reals poden representar-se sovint per mitjà de grans xarxes (grafs) i una de les característiques més rellevants d'eixes xarxes, és l'estructura de comunitats. Atés que la sociologia, la biologia o les situacions clíniques, poden modelar-se usualment per mitjà de grafs, els algoritmes de detecció de comunitats s'estan estenent cada vegada més en l'àmbit biomèdic.

En la present tesi doctoral, s'han fet contribucions en els tres camps anteriorment mencionats. D'una banda, s'han utilitzat mètodes d'avaluació de variabilitat temporal i multi-font, basats en geometria de la informació, per a detectar la variabilitat en la distribució de les dades que puga dificultar la reutilització dels mateixos i, per tant, les conclusions que es puguen extraure. Esta metodologia va demostrar ser útil després de ser aplicada als registres electrònics sanitaris d'un hospital al llarg de 7 anys, on es van detectar diverses anomalies.

A més, es va demostrar l'impacte positiu que esta anàlisi podria afegir a qualsevol estudi. Per a això, en primer lloc, es van utilitzar tècniques d'aprenentatge automàtic per a extraure les característiques més rellevants, a l'hora de classificar la intensitat del mal de cap en pacients amb migranya. Una de les propietats dels algoritmes d'aprenentatge automàtic és la seua

capacitat d'adaptació a les dades d'entrenament, en bases de dades en què el nombre d'observacions és xicotet, l'estimador pot estar esbiaixat per la mostra d'entrenament. La variabilitat observada després de la utilització de la metodologia, i considerant com a fonts els registres dels pacients amb diferent intensitat del dolor, va servir com a evidència de la veracitat de les característiques extretes. En segon lloc, es va aplicar per a mesurar la variabilitat entre els histogrames dels nivells de gris de mamografies digitals. Es va demostrar que esta variabilitat estava produïda pel dispositiu d'adquisició i després de la definició d'un preprocés d'imatge, es va millorar el rendiment d'un model d'aprenentatge profund, capaç d'estimar un marcador d'imatge del risc de desenrotllar càncer de mama.

Donada una base de dades que arreplegava les respostes d'una enquesta formada per escales psicomètriques, o el que és el mateix qüestionaris que servixen per a mesurar un factor psicològic, com ara depressió, resiliència, etc., es van definir noves arquitectures d'aprenentatge profund utilitzant l'estructura de les dades. En primer lloc, es dissenyà una arquitectura, utilitzant l'estructura conceptual de les esmentades escales psicomètriques. La dita arquitectura, que tractava de modelar el grau de felicitat dels participants, després de ser entrenada, va millorar la precisió en comparació amb altres models basats en estadística clàssica. Una segona aproximació, en la que l'arquitectura es va dissenyar de manera automàtica emprant detecció de comunitats en grafs, no sols va ser una contribució de per si per l'automatització del procés, sinó que, a més, va obtindre resultats comparables a la seua predecessora.

1 Introduction

Life need not be easy, provided only
that it is not empty.

Lise Meitner.

1.1 Motivation

Since Wang and Strong Data Quality (DQ) definition that consisted of 15 data quality dimensions classified in four high-level categories -*Intrinsic, Contextual, Representational* and *Accessibility* [1], numerous data quality frameworks have emerged and applied to a wide range of scenarios [2–5]. The complexity of DQ demonstrated by Rajan et al. [6] by showing the ongoing controversy flying on the concept. It is probably due to the blurry frontiers between the concepts included in each dimension definition. In any case, the common key is the idea of a DQ assessment is imperative to assure the data fitness-to-use. In this sense, DQ assessment methods emerge as an artifact helping in the community acceptance of the results of research works [7], and standardizing these methods could lead the transparency and consistency of DQ concept [8].

The controversy around DQ dimensions and concepts is also notorious in the biomedical field. Since Kahn et al. proposed [3] to classified DQ dimensions in two DQ main concepts -*Intrinsic* and *Conceptual*-, many approaches have been proposed in a biomedical environment. Weiskopf et al. highlighted five exhaustive and mutually exclusive dimensions [9] after a comprehensive literature review, and classified each found concepts into one of their proposed dimensions -*Completeness, Correctness, Concordance, Plausibility* and *Currency*. Almutiry et al. developed methods to measure DQ concepts *Accuracy, Consistency, Completeness*, and *Timeliness* [10]. The work of Johnson et al. [11] provides methods to assess enough DQ for secondary use of Electronic Health Records (EHR), it brings methods to measure the level of *Correctness, Consistency, Completeness* and *Currency*. The work of Kahn et al. [12] harmonized some of the EHR DQ frameworks into three categories *Conformance, Completeness,* and *Plausibility* and two frameworks to DQ assessment *Verification* and *Validation*. Finally, Rajan et al. [6] designed a computable data

quality knowledge repository for assessing quality and characterizing data in health repositories to standardize the data quality concepts and methods to assess them.

Clinical research may benefit from the power of the new artificial intelligence based technologies. Large multicenter and longitudinal studies could improve and generalize conclusions by using such techniques. But its application needs to ensure data is comparable among centers and different timespaces. These concepts could be classified in *Concordance* and *Timeliness* DQ dimensions.

Arthur Lee Samuel provided the first Machine Learning (ML) definition in 1959: *"Machine Learning gives computers the ability to learn without being explicitly programmed"*. From a mathematical point of view, Tom Mitchell's definition in 1977 deserves recognition: *"A computer program is said to learn from experience E concerning some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E"*. It must note that ML nowadays offers an operational definition of the problems rather than a cognitive definition. It means that a machine can learn to solve a problem intelligently through the experience instead of learning to think about how the problem can be solved, as was proposed by Alan Turing [13]. In this work, the question *"Can a machine think?"* replaced with *"Can a machine do what we (as thinking entities) can do?"*.

This thesis copes with the concepts mentioned above in the healthcare medicine field. It focuses on characterizing the variability among different sources and time batches in heterogeneous healthcare databases and assessing how the knowledge of this variability may lead to the improvement of ML-based models' performance, in particular, *Deep Learning* (DL) models. Furthermore, bearing in mind the need for the models to be deployed in healthcare scenarios, probably without data scientist supervision, the automation of the data-driven model design was also evaluated. These concepts established the main goals of this thesis, leading to the following research questions and objectives.

1.2 Research questions and objectives

The clinical conclusions extracted from local research studies could be universalized by carrying large longitudinal studies or studies covering many hospitals. Hospital management, population changes, or epidemics may impact data records deriving in a lack of quality for its reuse. In this sense, it becomes crucial to assess the temporal and multisource variability before data reuse. Besides, one of the ML applications is to extract variables influencing a determined outcome. If we consider a classification problem and measure the variability of inter sources -understanding as sources the classes- not to have a significant variability between classes would not give us any evidence of the reliability of the extracted discriminative variables.

ML is a subset of the methods included in the Artificial Intelligence concept. These algorithms have demonstrated good performance in many fields, including healthcare. Far from being the panacea, the application of such algorithms requires deep knowledge of themselves and the field of the study. The research of how these models could take advantage of the conceptual structure (for structured data) or observed data (for unstructured data) may be one of the first steps for model design automation based on machine learning. It could also lead to a standardization of the models to be deployed in different environments with the same target.

Healthcare medicine is a source of a great amount of data and establishes a wide variety of challenges. This thesis focuses on assessing temporal and multisource variability (understood as a DQ problem) and how the assessment of such DQ concepts may influence data reuse and methods to data-driven model design. The research questions that motivated the current document are:

RQ1 - Can data capture the complexity of a hospital working? Are temporal and multisource variability assessment methods capable of serving as a monitoring system of the DQ for its reuse?

RQ2 - May ML models use the information extracted from the application of temporal and multisource variability assessment methods to give credibility to factors influencing a determined outcome or increase performance?

RQ3 - Would it be possible to automate the process of DL model design using the conceptual structure of data? And for unstructured data, could the architecture design be generalized for other data types?

RQ4 - In the health field, it is imperative to be able to measure the influence of the variables in the prediction. Can this automatic architecture provide methods to evaluate the influence of variables on prediction?

The research work conducted in this thesis aims to provide answers to these questions. Theoretical scientific methods have been applied to a wide range of scenarios. DQ assessment methods were applied to characterize the temporal evolution of the EHR of a hospital, the intensity pain difference between headache women, and the difference of the histograms of mammographies acquired with different devices. The design of DL architectures was carried out, using a case of use, the estimation of the happiness degree (understanding happiness as a psychological factor) through five psychometric scales. To this end, the following objectives were defined:

O1 - Review the state-of-the-art of DQ, especially in a biomedical environment. Given the dimension of such a concept, the focus was on two inherent-to-data features such as temporal and multisource variability.

O2 - Review the state-of-the-art of ML models in a clinical environment. It covered the use of such algorithms in the psychology field, in particular in studies using Likert scales, and the use of DL algorithms in medical image analysis.

O3 - Evaluate the suitability of a temporal variability assessment methods to extract information about time evolution with data reuse purposes.
O4 - Demonstrate that using a multisource variability assessment method may give credibility to the feature importance extracted by an ML model and may help improve the performance of those models.
O5 - Develop a DL framework to automatically design the architecture for a specific type of structured data using its conceptual structure, which also allows us to measure the importance of the variables in the outcome.
O6 - Generalize the previous framework to be used with other data types. It relies on the building of the architecture by using the observed data instead of the conceptual structure.

The aforementioned objectives enclose the main goal of this thesis: *the study of how machine learning algorithms can take advantage of information captured in data.* Such a goal has been approached from two scopes: Firstly, *data quality (the temporal and multisource variability)*, which aims to capture differences in data distributions before model design, and secondly, *machine learning*, which aims to automatically build data-driven deep learning architectures that could be used without any prior knowledge on data. The following scientific contributions support the achievement of the proposed objectives.

1.3 Thesis contributions

This section presents the main contributions of this thesis. First, a summary of the most relevant aspects of each contribution is shown. Next, the scientific publications in high impact factor journals and conferences are listed.

1.3.1 Main contributions

C1 - Analysis of the temporal variability of data in the EHR of a hospital over seven years.
In this study, a temporal variability assessment method called TVA, based on information geometry, was applied to the EHR of a hospital. Data was composed of information about hospital admissions over seven years and the comorbidities of each patient. We were able to find some evidence on data distributions that explained (1) a hospital relocation, (2) a services reconfiguration, (3) a care-services redistribution, (4) the assignment to the hospital of a new area covering 80000 more patients, and (5) a pre-surgery admission protocol change. Data distribution changes were motivated by management decisions, so to a certain extent, they were expected. Probably what it was not expected was the scope of such changes since many times these decisions influenced unexpected factors. This work was published in the journal contribution **P1**.

C2 - The use of multisource variability assessment methods with feature extraction and prediction purposes.
A multisource variability assessment method called MSV, also based in information geometry, was applied to two databases. This contribution can be separated in the following "sub-contributions":

C2.1 - Study of the reliability of the feature importance extracted using a ML algorithm.
For this work, a database composed of clinical features and physical examinations from 67 migraine women was considered. It was tried to extract features influencing both intensity and frequency headache. Random Forest is an ML algorithm based on decision trees, which highlight by its ability to the adaptation to the decision space. With so few samples, it is easy to obtain models over-fitted to the training data. The application of the MSV assessment method allowed us to have scientific evidence that the features that Random Forest discovered as important when headache intensity estimated were certainly influencing. Meanwhile, the MSV assessment method applied to the headache frequency did not give any evidence of the feature importance was concluding. This work was published in the journal contribution **P2**.

C2.2 - The improvement of a deep learning model performance estimating the dense tissue in breasts.
The breasts are majority composed of *fatty tissue*. The appearance of *fibroglandular (or dense) tissue* is known to be an important biomarker of the risk of developing breast cancer. This study contained the full-field digital mammographies from 1785 women from eleven screening programs and its dense tissue segmentation made by two radiologists. After applying the MSV assessment method, we realized that there existed important differences in the image histograms from mammograms acquired using different devices. This work proposes a DL model to make a parametric dense tissue segmentation in breasts where the performance was significantly increased by using a simple preprocessing step standardizing the histograms of the images. The details of the preprocessing step, DL architecture and results were published in the journal contribution **P3**.

C3 - A method to design DL architectures using the conceptual data structure.
The *Data-Structure driven architecture for Deep Neural Networks (D-SDNN)* is a method to design DL architectures using the conceptual structure of (structured) data. The Likert scales are point scales that are used to allow the individual to express how much they agree or disagree with a particular statement. Its use is widely extended in social/behavioral sciences. These scales usually measure a factor which, in turn, is divided into sub-factors that are measured by a subset of items

of the scale. This structure was used to design a happiness/depression degree predictor using socio-demographic data and the responses to five Likert scales measuring five psychological factors. The training of such a model provided better results than state-of-the-art methods in psychology. Besides, two metrics to measure the importance of each item in the prediction are proposed. The results obtained in this work were included in the journal contribution **P4**.

C4 - A methodology to automatically design DL architectures using Community Detection algorithms in large networks.
A framework to automatically design DL networks, named *Community Detection based Deep Neural Network (CD-DNN)*, is proposed. It is based on the construction of a graph from a relationship between model inputs. Once the graph is built, a community detection algorithm at different resolutions is applied to infer the structure of data automatically. This framework has been developed to work with structured data, in particular Likert scales, but it would be easily extended to other data types. The metrics mentioned in Contribution **C3** are also applicable to models designed using CD-DNN. The proposal of the methodology was presented in the conference contribution **P5** and the journal contribution **P6**. The performance of the model was compared to the results obtained with the architecture of the previous contribution. The CD-DNN architecture obtained better results than those obtained by D-SDNN, besides the automation of the model building is an added value. These results were also published in the journal contribution **P6** and presented in the conference contribution **P7**.

1.3.2 Scientific Publications

The scientific contributions of this thesis have been published in four scientific top-ranked journals and two conference proceedings in the fields of Multidisciplinary Sciences, Clinical Neurology, Computer Science, Biomedical Engineering, Medical Informatics, and Applied Mathematics. All the following journal papers are either published or accepted for publication in journals included in the Journal Citation Reports (JCR). The versions presented in this dissertation are adaptations for the thesis due to university regulations. Each of them as a chapter having the same structure and bibliography as the original published version:

P1 - **Francisco Javier Pérez-Benito**, Carlos Sáez, José Alberto Conejero, Salvador Tortajada, Bernardo Valdivieso, Juan Miguel García-Gómez. "Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years". *PLoS ONE* (2019) 14(8) e0220369.

IF: 2.740 (JCR 2019): 27/71 (Q2) Multidisciplinary sciences.

P2 - **Francisco Javier Pérez-Benito**, José Alberto Conejero, Carlos Sáez, Juan Miguel García-Gómez, Esperanza Navarro-Pardo, Lidiane Lima Florencio, César Fernández-de-las-Peñas. "Subgrouping factors influencing Migraine intensity: A semi-automatic methodology based on machine learning and information geometry". *Pain Pract.* (2020) 20(3) 297-309.

IF: 2.258 (JCR 2019): 20/32 (Q3) Anesthesiology, 125/204 (Q3) Clinical Neurology.

P3 - **Francisco Javier Pérez-Benito**, François Signol, Juan-Carlos Perez-Cortes, Alejandro Fuster-Baggetto, Marina Pollán, Beatriz Pérez-Gómez, Dolores Salas-Trejo, María Casals, Inmaculada Martínez, Rafael Llobet. "A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation". *Comput. Meth. Prog. Biomed.* 195 (2020) 105668.

IF: 3.632 (JCR 2019): 30/109 (Q2) Computer science, interdisciplinary applications, 16/108 (Q1) Computer science, theory & methods, 22/87 (Q2) Engineering, biomedical, 6/27 (Q1) Medical informatics.

P4 - **Francisco Javier Pérez-Benito**, Patricia Villacampa-Fernández, José Alberto Conejero, Juan Miguel Garcá-Gómez, Esperanza Navarro-Pardo. "A happiness degree predictor using the conceptual data structure for deep learning architectures". *Comput. Meth. Prog. Biomed.* 168 (2019) 59-68.

IF: 3.632 (JCR 2019): 30/109 (Q2) Computer science, interdisciplinary applications, 16/108 (Q1) Computer science, theory & methods, 22/87 (Q2) Engineering, biomedical, 6/27 (Q1) Medical informatics.

P5 - **Francisco Javier Pérez-Benito**, E. Navarro Pardo, Juan Miguel García-Gómez, José Alberto Conejero. "Network clustering strategies for setting happiness degree predictors based on deep learning architectures". *Modelling for Engineering and Human Behaviour 2018.* Instituto de Matemática Multidisciplinar, Universitat Politècnica de València. Valencia, Spain. July 2018.

P6 - **Francisco Javier Pérez-Benito**, Juan Miguel García-Gómez, Esperanza Navarro-Pardo, José Alberto Conejero. "Community detection based deep neural network architectures: a fully automated framework based on Likert-scale data". *Math. Method. Appl. Sci.* 43 (2020) 8290-8301.

IF: 1.626 (JCR 2019): 67/260 (Q2) Mathematics, applied.

P7 - **Francisco Javier Pérez-Benito**, José Alberto Conejero, Juan Miguel García-Gómez, Esperanza Navarro-Pardo. "Community detection based architectures for deep learning: a fully automated framework for Likert scales".9^{th} *International Congress on Industrial and Applied Mathematics (ICIAM 2019).* International Council for Industrial and Applied

Mathematics and Sociedad Española de Matemática Aplicada. Valencia, Spain. July 2019.

1.4 Projects and partners

During the development of this thesis, the author has actively participated in research projects:

DQV-MINECO *Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos.* Funded by the Spanish Ministry of Economy and Competitiveness (Retos-Colaboración 2013 Programme, RTC-2014-1530-1, 2013-2016).

Objectives: This project aims to define a data quality evaluation and rating service to assure the data value aimed at its reuse in clinical, strategic, and scientific decision making. It will base on two software services. The first will evaluate nine data quality dimensions. The second will generate a data quality rating positioning the evaluated datasets according to reuse knowledge extraction purposes. These objectives fit the thesis objectives **O1**, **O3**, and **O4**.

Partners: VeraTech for Health S.L. (Valencia, Spain) and IBIME-ITACA group of the Universitat Politècnica de València (Spain).

ANÁLISIS DE CALIDAD Y VARIABILIDAD DE DATOS MÉDICOS. Funded by the Universitat Politècnica de València (Sept 2017-Feb 2019).

Objectives: This project has two key objectives. The first will be to assess the temporal and multisource variability of medical repositories. The second will evaluate how the heterogeneity of medical datasets may improve the performance of machine learning models. These objectives fit the thesis objectives **O1**, **O3**, **O4**, and **O6**. The framework of this project let the publication of the scientific contributions **P1**, and **P7**.

Partners: Hospital la Fe (Valencia, Spain) and Universitat Politècnica de València.

GVA18_HELPSALUD2 *Investigación aplicada a tareas reales en el sector de salud mediante técnicas de Machine Learning desplegadas en plataforma de Big Data Analytics.* Funded by Instituto Valenciano de Competitividad Empresarial (IVACE) (IMAMCN/2018/1 - HELPSALUD2, 2018-2019).

Objectives: This project aims to promote personalized health management as well as personalized prevention and diagnosis. With this purpose, it searched to develop healthcare technology that eases a larger interaction between specialists and users. These objectives fit the thesis contributions

O2, **O5**, and **O6**. In the framework of this project, the scientific publication **P3** published.

Partners: Instituto Tecnológico de la Infórmatica (Valencia, Spain).

1.5 Outline

This thesis is structured in eight chapters describing the research work carried out during the stage of development of this dissertation. Chapter 1 has introduced the motivation of this thesis, the research objectives, and the main contributions. Chapter 2 presents the theoretical background needed to complement the description of the methods developed in the compendium of articles included in this thesis. Chapter 3 describes the temporal analysis conducted in a hospital to identify unusual patterns through data. Chapter 4 analyses the differences between women with migraine; these differences lead to a different degree of intensity pain. Chapter 5 demonstrates how multi-source variability assessment methods may be used to improve deep learning models performance; the case of use is the parametric segmentation of dense tissue in digital mammographies. Chapter 6 proposes the D-SDNN method to build a deep learning architecture based on the conceptual data-structure and two metrics to measure the importance of the variables in the estimation, and it is applied to model a happiness/depression degree predictor. Chapter 7 automates the architecture design of the previous chapter by applying network clustering strategies after the definition of a model inputs relationship. Finally, Chapter 8 presents the concluding remarks and proposes recommendations to continue with the research developed in this thesis.

Figure 1.1 outlines the thesis contributions structured among the thesis chapters along with the publications developed during this study.

Contributions		Chapters	Publications	Main projects
		1. Introduction		
		2. Rationale		
C1 – Analysis of the temporal variability of data in the EHR of a hospital over seven years.		3. Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years.	**P1 – PLOS One**	**DQV-MINECO:** Servicio de evaluación y rating de la calidad de repositorios de datos biomédicos. (RTC-2014-1530-1)
C2 – The use of multisource variability assessment methods with feature extraction and prediction purposes.	**C2.1** – Study of the reliability of feature importance extracted using a Machine Learning algorithm.	4. Subgrouping factors influencing migraine intensity: A semi-automatic methodology based on machine learning and information geometry	**P2 – Pain Practice**	**GVA18-HELPSALUD2:** Investigación aplicada a tareas reales en el sector salud mediante técnicas de Machine Learning desplegadas en plataforma de Big Data Analytics. (IMAMCN/2018/1)
	C2.2 – The improvement of a deep learning model performance estimating de dense tissue segmentation.	5. A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation.	**P3 – Computer Methods and Programs in Biomedicine.**	
C3 – A method to design deep learning architectures using the conceptual data structure.		6. A happiness degree predictor using the conceptual data structure for deep learning architectures.	**P4 – Computer Methods and Programs in Biomedicine.**	
C4 – A methodology to automatically design DL architectures using Community Detection Algorithms in large networks.		7. Community detection based Deep neural network (CD-DNN) architectures: a fully automated framework based on Likert-scale data.	**P6 – Modelling for Engineering and Human Behaviour 2018.** **P7 – Mathematical Methods in the Applied Sciences.** **P8 – 9th International Congress on Industrial and Applied Mathematics.**	**Análisis de Calidad y variabilidad de datos medicos.**
		8. Concluding remarks & recommendations.		

Fig. 1.1: Outline of the thesis contributions. chapters, publications and projects.

2 Rationale

A bottle of wine contains more philosophy
than all books in the world.

Louis Pasteur.

The research methodology followed in this thesis was a *design science* approach [14]. By definition, this approach consists of taking existing technologies to generate knowledge and using this knowledge to improve conclusions. The technologies used in the development of the works that make up this thesis covered several fields. Methodologies based on *Information Geometry* were applied to assess the *Quality* of databases, *Machine Learning* algorithms -in particular, *Random Forest* and *deep learning*- were developed to solve some observational problems, and finally, *community detection* algorithms in *large networks* were applied to propose ways to automatically design machine learning-based models.

Since this document covers a wide range of technical disciplines, this chapter aims to give insights into each of the previously mentioned technologies to help readers follow the rest of the chapters. Each of the scientific papers enclosed (from Chapter 3 to 7) in this dissertation extends the information given in the current chapter.

As a brief, Chapter 3 contains the application of a DQ method based on Information Geometry to assess the temporal variability of a hospital EHR over seven years. In Chapter 4, we describe how this DQ method can serve as an artifact to measure the differences among the intensities of headache pain on migraine patients, and how these differences may reinforce the conclusions extracted from a machine learning-based model (Random Forest). In Chapter 3, we show a convolutional neural network (deep learning algorithm) that estimates two parameters to segment a specific area of breasts. Its performance has been improved after detecting differences among the histograms of mammograms by using the previously mentioned DQ method and applying a simple preprocessing to the image. Chapter 6 proposes a deep learning

architecture based on the conceptual data structure of a psychological survey to predict the happiness/depression level. Finally, Chapter 7 tries to solve the same problem of the previous chapter, but in this work, a framework based on community detection on graphs is proposed to design the deep learning architecture automatically.

2.1 Data Quality. Multisource and Temporal Variability Assessment Methods

The comparison of two probability distributions is defined as follows. Suppose a variable X follows a probability distribution $p(x)$. A measure of information from an observation x of $p(x)$ was proposed by Shannon [15] as:

$$f(x) = -\log\left(p(x)\right)$$

then the expected information in X, (the Shannon entropy) is given by:

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \text{ if } X \text{ is discrete}$$
$$H(X) = -\int p(x) \log p(x) \text{ if } X \text{ is continuous.}$$

From this entropy, the Kullback-Leibler divergence, which measures the information inefficiency of assuming a distribution Q when a true distribution P, is given by:

$$KL(P||Q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

This divergence is not a true distance, because it is not symmetric nor satisfies the triangle inequality. What drives to consider:

$$JSD(P,Q) = \left(\frac{KL(P||M)}{2} + \frac{KL(Q||M)}{2}\right)^{\frac{1}{2}}$$

where $M = \frac{1}{2}(P+Q)$, is known as Jensen-Shannon distance, and besides being a distance, it is also bounded by 0 and 1.

Information geometry is a field which translates the concepts and properties of differential geometry into spaces of probability distributions [16]. These spaces are known as *statistical manifolds* and lie in Riemannian spaces.

From a set of n non-parametric probability functions $(P_i, \ldots, P_n)$, the $\binom{n}{2}$ pairwise Jensen-Shannon distances may be computed. Let us define the dissimilarity matrix $Y = (y_{11}, \ldots, y_{nn})$ where $y_{i,j} = JSD(P_i, P_j)$. The objective of Multidimensional Scaling (MDS), which is a manifold learning algorithm

that allow us to obtain an Euclidean approximation of such spaces, is to obtain the set $E = (e_{11}, \ldots, e_{nc})$ of points in a Euclidean space $\mathbb{R}^c$, with $c \leq n-1$, in order to find the best approximation of $\|e_i - e_j\| \approx f(y_{ij})$, where $\|\cdot\|$ is the euclidean norm, and $f(\cdot)$ is a transformation of the original dissimilarities, optimally $f(y_{ij}) = y_{ij}$. This approximation can be solved by the minimization of the raw loss function:

$$\min_E \sum_{i<j} (f(y_{ij}) - \|p_i - p_j\|)^2,$$

After applying MDS to the dissimilarity matrix of the distances between the n probabilistic functions $(P_1, \ldots, P_n)$, a $n-(by)-c$ euclidean coordinates matrix $V = (V_{11}, \ldots, V_{nc})$ is obtained. Each row, $V_r = (V_{r_1}, \ldots, V_{rc})$, of this matrix represents the c coordinates in the euclidean $\mathbb{R}^c$ space of the r^{th} Probability Density Functions (PDF) P_r.

Therefore, a $c-$dimensional irregular simplex (if the distance between all pair of vertexes is equal, the simplex is said to be regular) S may be defined as:

$$S^c = (V, C),$$

where V corresponds to the coordinates of the vertexes and C to the simplex centroid, see(2.1).

$$C = \sum_{l=1}^{n} \frac{V_l}{n}, \tag{2.1}$$

Given a database, we can consider different groups of observations (either in terms of sources or temporal batches) and compute their multidimensional probability density functions (using for instance Kernel Density Estimation (KDE) [17]). After applying the previously mentioned methodology, we obtain the euclidean simplex $S = (V, C)$ where the statistical manifold defined by the PDFs is embedded.

The centroid C of the simplex $S = (V, C)$ may be understood as the latent central tendency of the original database. Then, the distance of a vertex $V_i \in V$ to C, $d(V_i, C)$, where $d(\cdot, \cdot)$ is the euclidean distance, represents the deviation of the group modeled and represented by V_i to the central tendency of the database.

If the PDFs (represented by V when simplex built) are considered as individuals of a population and the centroid of the simplex (C) the central tendency of the population, the standard deviation among n PDFs can be defined as:

$$Std(P_1, \ldots, P_n) = \frac{\sum_{i=i}^{n} d(V_i, C)}{n},$$

It should be noted that although the Jensen-Shannon distance is $[0, 1]$-bounded, the MDS transformation makes the distance between the pairs of

points in the vertex could be higher than 1, depending on the transformation of the space needed to embed the statistical manifold into the euclidean space. It makes this standard deviation not to be comparable among different studies. Besides, if a new group modeled by other probability density function is wanted to be added to the study, it will be mandatory to compute the simplex again [18].

A regular simplex fulfills that the distance from each of its vertexes to the centroid is equal. The angle between any segments joining a vertex with the centroid depends on the simplex dimension (c), and is [19]:

$$\gamma(c) = \arccos\left(\frac{-1}{c}\right)$$

The $1-$regular (1R) simplex verifies that the distance any pair of its vertexes is one. It can be proven that the distance between any vertex of the $1R-$simplex to the centroid depends on the dimension and is given by:

$$d_{1R}(c) = \frac{1}{2\sin\left(\frac{\gamma(c)}{2}\right)},$$

where $d_{1R}(1) = \frac{1}{2}$. If we consider an irregular simplex defined as a simplicial space upper-bounded by a $1R$ simplex, then the distance of any vertex to the centroid of the irregular simplex will be bounded by:

$$d_{max}(c) = 1 - \frac{1}{c+1},$$

which is larger than $d_{1R}(c)$ for the same c [4].

Let us define two metrics [18], *Global Probabilistic Deviation* (GPD) and *Source Probabilistic outlyingness* (SPO), based on the simplex previously introduced:

- The global probabilistic deviation metric GPD among a set of groups of observations $X = (X_1, ..., X_n)$ is defined as:

  $$GPD(X_1, ..., X_n) = \frac{Std(P_1, \ldots, P_n)}{d_{1R}(c)},$$

  where P_i is the PDF for the group of observations X_i and c is the dimension of the euclidean simplex.
- The source probabilistic outlyingness metric (SPO) of a group of observations X_i with respect to the central tendency among the set of groups of observations $X = (X_1, ..., X_n)$ is defined as:

  $$SPO(X_i) = \frac{d(V_i, C)}{d_{max}(c)},$$

  where V_i are the euclidean coordinates for the $i_{th}-$vertex of the simplex, C is the centroid of the simplex, and c is its dimension.

2.1.1 Variability visualization methods

The methodology previously explained serves as a way to measure differences in the probability distributions of the variables among different groups of observations. A database given can also split by time batches (e.g., months, weeks, or years), and these methods provide measurements of the differences of data among these time batches. Some visualization methods based on the simplex and probabilistic distances provide a graphical way to assess how data evolve across time. We refer readers to the work of Sáez et al. [5] for further details:

- **Information Geometry Temporal (IGT) Plot.** It is a visualization of the simplex obtained after the MDS application. The simplex is projected into the two most relevant dimensions (D1-simplex and D2-simplex), which allow representing the dissimilarity over temporal batches as a two-dimensional (2D) plot. Each point of the plot is labeled to show the batch (date) that the point represents, furthermore the points are colored to facilitate the interpretation (cool colors for winter and warm colors for summer), and a smoothed timeline path is also provided. Thus, this visualization method helps in showing temporal trends in data, abrupt changes (high distance between adjacent time points), recurrent changes (recursive flow through specific areas), related periods, and anomalies (outlying points).
- **Probability Distribution Function Statistical Process Control (PDF-SPC) algorithm:** It is an adaptation of classical statistical process control (SPC) algorithms. It is applied to monitor the variability of data distribution through consecutive temporal batches. An upper confidence interval of the accumulated distances of temporal batches to a moving reference distribution (initially the first batch) is computed. The degree of change of the current time batch to the reference distribution is classified according to the magnitude of the current confidence interval. The possible states are: *in-control* (distribution are stable), *warning* (distribution are changing), and *out-of-control* (abrupt change detected with respect to the reference distribution). When a *out-of-control* state is reached, the reference distribution is set to the current. The PDF-SPC consists of plotting a control chart in which we represent the time batches on the $X-$axis and the distances on the $Y-$axis. The current distance to the reference, the mean of the accumulated distances, and the upper confidence interval are plotted for each time batch. If a *warning* state is observed, a vertical broken line is drawn, and if a *out-of-control* state is detected, a continuous vertical line is drawn.
- **Temporal heat map:** By representing temporal batches on X axis and a value (or range of values) on the Y axis, the color at the pixel (x, y) indicates the frequency at which value y was observed on date x. This approach also allows the monitoring of how the values of a variable evolve.

Similarly, a database can be split according to any feature (different to time, e.g. the origin of data or the hospital service where data was acquired). A visualization method based on the simplex and the SPO metric provides a way to assess the differences among any groups of observations [4].

- **Multi-Source Variability (MSV) Plot:** Just like on IGT-plot, the simplex is 2D projected. In this visualization, each point is represented by a circle whose radius is proportional to the number of cases in the group of observations it represents. The color of the circle indicates the SPO metric value.

This methodology is used in Chapter 3 to measure the temporal variability on EHR data from a hospital, in Chapter 4 to characterize differences among migraine patients with different headache pain intensity and frequency, and finally, in Chapter 5 to study differences among mammogram histograms.

2.2 Machine Learning

Every ML model is at least composed of two steps, training and validation. These models usually consist of a set of parameters that are updated during the training step to fit the problem to be solved. The training step is iterative, and the performance is computed in each iteration to do the parameter update until a tolerance value is reached. The validation step is typically a set of samples in which the response is known. This response is used to estimate the error of the model on the prediction of unseen samples. By monitoring the validation error, it is possible to avoid problems of particularization, which provoke the fail to predict future observations reliably. Such errors are known as overfitting.

The task of an ML model may be categorized according to the desired output as:

- **Classification:** The inputs of the models are labeled into two or more classes. After the training of the model, this can assign unseen inputs to one of these classes.
- **Regression:** The output, in this case, is continuous instead of discrete.
- **Clustering:** A set of inputs must be divided into a priori-unknown group.

According to how the ML model learns, the algorithms may be divided into two major categories:

- **Supervised learning:** This kind of algorithm needs the inputs to be labeled or have a continuous response. The model learns the way to best infer, according to a predefined performance score, the expected classification/response.

- **Unsupervised learning:** This type of algorithm learns data patterns on its own from observations without any associated response.

All the implemented ML models for this thesis contributions were trained using supervised learning. Classification and regression tasks were performed using different ML artifacts such as Random Forests (RF), and Deep Learning (DL).

Without going into greater detail, RF can provide information about what variables are the most important in the prediction task taking into account the training samples. Although these features' importance may decrease by voting share between the correlated predictor variables, RF is a good option because of its accuracy, robustness, ease of use, and relative interpretability. Meanwhile, DL is probably the most powerful artifact in terms of performance. Such has been demonstrated in many works of different fields [20–22]. But these DL models are usually applied in a black-box manner, and the lack of interpretability can mean a major drawback in medical applications [23]. It has led to the emergence of the *Explainable Artificial Intelligence (EAI)* research line. Contribution 3 aligns with the EAI topic. It tries to exploiting the DL powerful performance but proposing a way to interpret the importance of the variables in the outcome.

Let us give some insights into each of the aforementioned technologies.

2.2.1 Ensemble Learning

Ensemble learning is a set of methods that use multiple algorithms to obtain a better predictive performance than those that could get from each of the algorithms alone. During the development of this thesis, we used an ensemble learning algorithm known as Random Forest (RF). We can understand an RF as a strong predictor that makes a prediction based on the pooled results of several weak predictors, and for RF, these weak predictors are Decision Trees (DT).

2.2.1.1 Decision trees

A DT may be considered as a directed graph $G = (V, E)$ in which any two vertexes are connected by exactly one path, so all the edges are directed away from the root. If there exists an edge from v_1 and v_2 (*i.e.*, if $(v_1, v_2) \in E$), then v_1 is said to be the *parent* of v_2 while v_2 is said to be a *child* of v_1. A node with no child is known as *leaf*.

Any regression or classification task $f : X \longrightarrow Y$, may be represented by a DT. The root represent the whole input space X, and each node v_i represents a subspace of the space represented by its parent $X_i \subseteq X_p$ where $(v_p, v_i) \in E$. The set of the children of a node v_p is $\{v_c \mid (v_p, v_c) \in E\}$ and it is verified that $\bigcup X_c = X_p$.

The way a node is split into child nodes is under a condition on one of the space features. This condition is chosen among all the possible conditions on

the whole set of features based on the concept of node impurity. The lower node impurity, the better partition of the parent set.

For regression tasks, common metrics to measure the impurity are the variance, taking the mean squared error (MSE) or the mean absolute error (MAE). For classification tasks, Gini impurity and Entropy are commonly used (see the Table 2.1):

Impurity	**Task**	**Formula**
Variance MSE	Regression	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$
Variance MAE	Regression	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \mu\rvert^2$
Gini	Classification	$\sum_{i=1}^{C} f_i(1 - f_i)$
Entropy	Classification	$\sum_{i=1}^{C} -f_i \log(f_i)$

Table 2.1: Impurity formulas. For regression, the number of elements in the set associated to a node is N, $\{y_i\}$ is the observed target values for the elements of a node. For classification, C are the set of classes, and f_i represents the frequency of the class i at a node.

These metrics are lower bounded by 0. For classification, both Gini and Entropy impurity reach a value of 0 if and only if all the elements belong to the same class. For regression, both MSE and MAE variance is 0 if and only if all the elements of the node have the same target value.

2.2.1.2 Random Forest

Let us introduce two concepts related to ensemble methods. Given a training set $T = \{x^1, \ldots, x^m\}$ with m observations of D features, -i.e. $x^i = (x_1^i, \ldots, x_D^i)$ for each $i \in \{1, \ldots, m\}$-:

- **Bootstrap aggregating (Bagging)** is a strategy that provides t new training sets $\{Ti\}_{i=1}^{t}$ of a predefined size $m' < m$. These new sets are uniformly sampled from T with replacement.
- **Random Subspace** $l-$dimensional of T provides a new training set (T') with m observations of $l < D$ features. These features are randomly sampled with replace what allows the repetition of features in the set.

Apply bagging we obtain t new training sets $\{T_1, \ldots, T_t\}$ with m' samples each set for training an RF model composed of t DTs. Then, we compute Random Subspace with a predefined number of features l to each of these new t sets. Finally, we train each of the DTs with one of the t new training sets of

size (m', l). If a regression task considered, the result of the RF is the mean or the median of the DTs' results, and if classification task, the resulting class after applying the RF will be the most voted class by the DTs.

This algorithm has many advantages. It can be highlighted that RF works well with categorical and continuous variables and can handle missing values. Besides, it does not need data to be standardized and usually is robust to outliers. Finally, it provides a way to compute features' importance.

2.2.1.3 Interpretability

Firstly, let us see how to compute the importance of each feature for a DT. Let v_j be the j^{th} node (space X_j), let us name $\{v_i^j \mid i \in H\}$ (spaces $\{X_i^j \mid i \in H\ ,\ X_i^j \in X_j\ ,\ X_l^j \bigcap_{l \neq k \in H} X_k^j = \emptyset\ ,\ \bigcup X_i^j = X_j\}$) the set of the children of the node j, and I_j the impurity of the node j after the model training. Then, the importance of the node j may be computed according to the following formula:

$$NI_j = w_j I_j - \sum_{i \in I} w_i^j I_i^j, \text{ where } w_l = \frac{\#X_l}{\#X}.$$

Let us j represent the nodes splitting by a condition on the feature l, the importance of the feature l is given by:

$$FI_l^{DT} = \frac{\sum_j NI_j}{\sum_{v_k \in V} NI_k}$$

Once the feature importance have been computed for each DT composing the RF, the i_{th}-feature importance for the RF model is computed by averaging its importance among the DTs:

$$FI_i^{RF} = \frac{1}{N} \sum_{j=1}^{N} FI_i^{DT_j}$$

where N is the number of DTs that compose the RF and DT_j represents the j_{th} DT.

Random Forest is used in Chapter 4 to model the headache intensity in migraine patients and to assess which variables are the most discriminatory to classify patients according to its pain intensity.

2.2.2 Deep Learning

Around 1900, David Hilbert presented a famous collection of problems that set the century mathematics research. The 13th problem considered the possibility of expressing a function of n variables as the combination of sums and compositions of two functions of a single variable.

Kolmogorov (1957), Arnold (1958), and Sprecher (1965) provided proofs that there must exist such representation. This is known as *Kolmogorov-Arnold representation theorem*, and its formulation is that shown in Equation 2.2.

$$f(x_1, \ldots, x_D) = \sum_{q=0}^{2D} \Phi \left(\sum_{p=1}^{D} \lambda_p \phi(x_p + \eta q) + q \right) \tag{2.2}$$

where η and the $\{\lambda_p\}$ are real numbers, and Φ and ϕ are univariate functions.

Without loss of generality, let us now consider a simple Neural Network with one hidden layer composed of three neurons (see Figure 2.1). This model is not considered to be a DL model, but its working is the same and it is a good choice to introduce the technology.

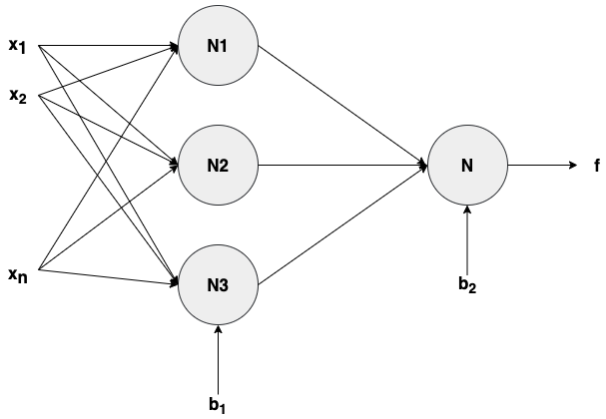


Fig. 2.1: An example of a simple 1-hidden-layers neural network.

The output of each neuron is the application of a function, called *activation function* f that in this case we supposed is the same for all the neurons, to the sum of the weighted sum of its inputs plus a number called *bias*. It is easy to conclude that:

$$f(x_1, \ldots, x_D) = f \left(\sum_{j=1}^{3} w_{j1}^{(2)} f \left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + b_{1i} \right) + b_2 \right) \tag{2.3}$$

where, w and b are the parameters of the model which are randomly initialized and updated at each training iteration using derivative information. The similarity between 2.2 and 2.3 equations is evident. In this sense, a neural network model may be understood as an "easy univariate representation" of a complex $D-$dimensional function. The training stage is a numerical method to fit the model parameters, w and b, to the observations.

The most typical activation functions are the following:

- **Linear function:** It ranges $[-\infty, \infty]$ is $0-$oriented and easy to compute.

$$f(x) = x \text{ whose derivative is } f'(x) = 1$$

- **Sigmoid/Logistic function:** It ranges $[0, 1]$ is not $0-$oriented and its computation is intensive.

$$f(x) = \frac{1}{1 + e^{-x}} \text{ whose derivative is } f'(x) = f(x)(1 - f(x))$$

- **Rectified Linear Unit (ReLu):** It ranges $[0, \infty]$ is not $0-$oriented and its easy to compute.

$$f(x) = \begin{cases} 0 \ for \ x < 0 \\ \\ x \ for \ x \geq 0 \end{cases} \text{ whose derivative is } f'(x) = \begin{cases} 0 \ for \ x < 0 \\ \\ 1 \ for \ x \geq 0 \end{cases}$$

- **Leaky ReLu:** It ranges $[-\infty, \infty]$ is not $0-$oriented but it could be, and its easy to compute.

$$f(x) = \begin{cases} \alpha x \ for \ x < 0 \\ \\ x \quad for \ x \geq 0 \end{cases} \text{ whose derivative is } f'(x) = \begin{cases} \alpha \ for \ x < 0 \\ \\ 1 \ for \ x \geq 0 \end{cases}$$

where $\alpha \in \mathbb{R} - \{0\}$.

The *loss function*, *cost*, is a function that quantifies the difference between the predicted values by the model and the expected ones. This function having in mind the task to be carried out. For each training iteration, from now on *epoch*, the sum of the loss function of all training samples is computed to obtain the estimation error.

Let it be $x = x^{i}{}_{i=1}^{m}$ the training samples and $y = y^{i}{}_{i=1}^{m}$ the observed values with $i \in [1, \ldots, D] \subset \mathbb{N}$. The estimation error E is given by:

$$E = cost(x, y)$$

This step is known as *forward step* and is the evaluation of the model over the set of training samples. After this, the *backpropagation step* begins, it consists of the minimization of the computed error E by updating the values of the model parameters $\{w\}$ and $\{b\}$ using the derivative information. Without losing generality, suppose that the activation function is the same for all the neurons in our model, f.

Let us denote:

$a^{(1)} = \{x_i\}_{i=1}^{D}$ is the input layer

$z_n^{(l)} = \sum_{i=1}^{D_l} w_{in} a_i^{l-1} + b_i^{l-1}$ is the n^{th} neuron value at hidden layer l.

$a^{(l)} = f\left(\sum_{j=1}^{n_l} z_j^{(l)}\right)$, where n_l is the number of neurons in the layer l.

Using the chain rule, is easy to conclude that:

$$\frac{\partial E}{\partial w_{in}^{(l)}} = \frac{\partial E}{\partial z_n^{(l)}} \frac{\partial z_n^{(l)}}{\partial w_{in}^{(l)}} = \frac{\partial E}{\partial z_n^{(l)}} a_i^{(l-1)}$$
$$\frac{\partial E}{\partial b_n^{(l)}} = \frac{\partial E}{\partial z_n^{(l)}} \frac{\partial z_n^{(l)}}{\partial b_n^{(l)}} = \frac{\partial E}{\partial z_n^{(l)}} 1$$

The *local gradient* $\frac{\partial E}{\partial z_n^{(l)}}$ can be easily computed by going backwards through the different layers. It is known as Gradient Descent (GD). The GD is applied over the cost function applied to the whole training set meanwhile Stochastic Gradient Descent (SGD) performs a parameter update for each training sample. Once all of them have been obtained, the model parameters are updated according to:

$$\begin{aligned} w_{in}^{(l)} &= w_{in}^{(l)} - \epsilon \frac{\partial E}{\partial w_{in}^{(l)}} \\ b_n^{(l)} &= b_n^{(l)} - \epsilon \frac{\partial E}{\partial b_n^{(l)}} \end{aligned} \tag{2.4}$$

where ϵ is the learning rate and determines the gradient's influence. It is also worth mentioning that several adaptative gradient descent algorithms have been proposed in which the learning rate is also updated at each stage. They have demonstrated a better convergence than unadaptative algorithms such as GD and SGD. Examples of these algorithms are *Adagrad* [24] and *Adam* [25]. We refer readers to the original works to extend the information about such algorithms.

Each forward and backpropagation step is a training epoch, and it is repeated until a predefined number of epochs or a convergence stop criterion reached.

These are the basics of how artificial neural networks learn to solve a problem. Through observations, they are capable of updating their weights to best fit the loss function wanted to minimize. A wide range of possibilities are now available, and they are being exploited. Although the literature is extensive, let us introduce some examples of modern architectures considered deep learning approaches.

If the architecture shown in Figure 2.1 is updated by adding some more hidden layers, the obtained model is known as *Deep Feedforward Neural Networks (DFFNN)* [26], *Recurrent Neural Networks* are those whose neurons of the hidden layers receives its output with a fixed delay [27] and by adding to each neuron two elements called *gates* that control previously processed information, neurons can be provided with memory which give rise to the *Long Short Term Memory* neural networks [28]. A new convolutional layer provides a way to extract information from images (e.g. edges, shapes, etc.) by applying filters. Deep Feedforward Neural Networks are frequently attached to the final convolution layer for further data processing. These architectures are known as *Convolutional Neural Networks* [29] and will be deeper introduced in the next section.

2.2.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are those architectures that use two operations called *convolution* and *pooling* to reduce an image to its essential features. Typically these features serve to classify the image (usually by the application of DFFNN after the convolution stages).

An image can be understood as a matrix in which each cell represents a pixel and contains a number identifying the color. For color images each cell typically contains a tuple (r, g, b) to describe the intensities of the three additive primary colors (channels) *Red*, *Green*, and *Blue* (RGB) which reproduce the color of the pixel. For gray-scale images, each cell only contains a number (one channel) identifying the gray intensity. The digital mammographies which were used in the development of this thesis were a gray-scale. In this sense, the next explications will be given under starting from a 1-channel image.

In computer vision, a convolution is an operation where a small matrix of numbers called *kernel* or *filter* is used to transform the values of the image by recursively applying it to consecutive image windows of the same size of the kernel, until the whole image covered. Formally, let us denote by I a 1-channel image and h a kernel, which without loss of generality may suppose square, of sizes (n_I, n_I) and (n_h, n_h) respectively.

Let us introduce two concepts needed to apply a convolution, *padding* and *strides*:

- *Padding (p):* is the addition of pixels (usually with a value of 0) around the image to control the output of the image after the application of the convolution. If we want to obtain an image with the same size as the input, p pixels should be added at each border:

$$p = \frac{n_h - 1}{2}, \text{ where } n_h \text{ is the kernel dimension.}$$

  Previous equation is under the assumption of having a stride of 1.

- *Stride (s):* is other concept to control the size of the output of the convolution operation. It refers to the number of pixels shifted when a new image window is selected to apply the kernel. To ease the understanding, we will suppose the horizontal and vertical strides are the same. The dimension of the convolution output matrix, with p padding and s stride, is given by:

$$n_{out} = 1 + \frac{n_I + 2p - n_h}{s}.$$

Once the previous concepts defined, the result of the convolution by the kernel h applied to the image f, with padding p and stride s, will be a matrix G of size (n_{out}, n_{out}) where the pixel of the row r and column c obtain the value:

$$G[m,n] = (f * h)[m,n] = \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} h[i,j] f[m-i, n-k]$$

It should be noted that many times, *pooling* layers are used to reduce the dimensions of an image after a convolution. The most typical pooling layer is the *maxpooling*, which covers the image using disjoint windows of size (n_m, n_m), the resulting matrix of size $(\frac{n_I}{n_m}, \frac{n_I}{n_m})$ will be composed of the maximum value obtained in each window.

A convolutional layer is composed of a set of convolutions to which an activation function f is applied after adding bias. Making the comparison with Equation 2.3 of section 2.2.2:

- The weights w are the kernels h. The cell values of each kernel will be updated during the training stage.
- The operation $(\sum)$ of each neuron of the conventional neural network is replaced by the convolution. The output of a neuron was a number, and CNN is a matrix.
- The bias for each convolutional layer will be a vector of matrices instead of a vector. This vector of matrices will be composed of as many matrices as kernels the convolution had. After making the sum of each convolution to its bias, the activation function is applied to each matrices element.

The training is similar to the artificial neural networks explained at the beginning of the section. The forward step goes transforming the image across convolutions. When the final of the model is reached, the loss function is evaluated and the derivative information is backpropagated, updating the kernel values to obtain the best images representation to solve the task in hand.

A deep architecture where not all the hidden layers were connected used to model a task in Chapter 6 and Chapter 7. Chapter 5 provides a variation of a convolutional neural network to learn breast segmentation parameters as intrinsic features of the image.

2.3 Graphs

In the domain of mathematics, physics, and computer science, *graph theory* is the study of graphs, which are structures to model relationships between individuals. This field has raised the interest of many great scientists, we refer readers to the Euler's *Seven bridges of Königsberg* [30] to understand a classical problem to be solved with graphs.

Formally, a graph is a pair $G = (V, E)$ where:

- $V = \{v_i\}_{i=1}^{n}$ is the set of vertexes.
- $E = \{(v_1, v_2) \in V \times V\}$ is the set of edges.

If the edges have not orientation, i.e. if $(v_1, v_2) \in E$ then v_2 can be reached by v_1 and v_1 reached by v_2, then the graph is said to be *undirected*. If the graph presents oriented edges, i.e. if $(v_1, v_2) \in E$ implies that v_2 can be reached by v_1 but not vice versa, then the graph is called *directed*.

Furthermore, the edges can have a weight associated, this weight is a way to quantify the relation between two vertexes. The graphs whose edges have a weight associated are known as *weighted graphs* and are represented as $G = (V, E, W)$, where $W = (w_{ij}) \mid i, j \in 1, \ldots, n$ is a $n \times n$-square matrix and a specific element of this matrix w_{ij} is the weight between the vertexes v_i and v_j. If $(v_i, v_j) \notin E$ then $w_{ij} = 0$.

An edge $e \in E$ is a *loop* if $e = (v, v)$ with $v \in V$, and two edges that have the same origin and end vertexes are said to be *parallel*. A graph is *simple* if it has no parallel nor loop edges and those graphs with only one vertex are denominated as *trivial graphs*.

The vertex $v_2 \in V$ is adjacent to $v_1 \in V$ if they are connected by an edge, i.e. $(v_1, v_2) \in V$. It is important to note that in undirected graphs v_1 and v_2 would be adjacent each other. Besides the edge $(v_1, v_2) \in E$ is said to be incident to vertexes v_1 and v_2 for undirected graphs, and incident to v_2 for directed graphs. It is important to note that two edges $e_1 = (v_{11}, v_{12})$, $e_2 = (v_{21}, v_{22}) \in E$ are *adjacent* if they share a common vertex:

- $v_{12} = v_{21}$ for directed graphs.
- $v_{1i} = v_{2j}$ where $i, j \in \{1, 2\}$ for undirected graphs.

The neighborhood of a vertex v, $N(v)$ is the vertexes adjacent to e, i.e. $N(v) = \{u \in V,$ where $(u, v) \in E$ for directed graphs, $(u, v) \in Eor(v, u) \in E$ for undirected graphs$\}$. Thus, the degree of a vertex v is defined as the number of vertexes in its neighborhood, $d(v) = \#N(v)$.

It said to exist a *path* between two vertexes $v_1, v_2 \in V$ if there is a set of edges joining v_1 to v_2, i.e. $p = \{(v_{1i}, v_{2i}) \in E$ with $i \in 1, \ldots, m, v_{11} = v_1, v_{2m} = v_2,$ and $v_{1(j+1)} = v_{2j}$ for $1 \in 1, \ldots, m-1$. The *distance* between two vertexes, $d(v_1, v_2)$, is the shortest path between both vertexes. If $\{p_k = \{(v_{lk_i}, v_{2k_i})\}\}$ is the set of paths between v_1 and v_2, then:

$$d(v_1, v_2) = \begin{cases} \min_k \#p_k & \text{for undirected graphs} \\ \\ \min_k \sum_{k_i=1}^{\#p_k} w_{1k_i 2k_i} & \text{for directed graphs} \end{cases}$$

The *eccentricity* of a vertex v, $\epsilon(v)$, is the maximum distance to any vertex from v, i.e. $\epsilon(v) = \max_{u \in V} d(v, u)$. While the *radius* of the graph G is the minimum eccentricity of any vertex, $r = \min_{v \in V} \epsilon(v) = \min_{v \in V} \max_{u \in V} d(v, u)$, the *diameter* of the graph G is the maximum eccentricity of any of its vertexes, $r = \max_{v \in V} \epsilon(v) = \max_{v \in V} \max_{u \in V} d(v, u)$. Finally, the *betweenness centrality of a vertex* v is:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma st},$$

where σ_{st} is the number of shortest paths from vertex s to vertex t, and $\sigma_{st}(v)$ is the number of shortest paths from vertex s to vertex t passing through v, and is a measurement of the importance of the vertex v inside the graph. And the *betweenness centrality of a edge* e is:

$$g(e) = \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma st},$$

where $\sigma_{st}(e)$ is the number of shortest paths from vertex s to vertex t.

Community detection

Given these notions on graphs, let us develop a brief on community detection in graphs, which will be extended in Chapter 7.

A community can be defined as a subset of vertexes that are densely connected to each other and loosely connected to other communities in the same graph. Recent literature uses this technology to solve a wide variety of problems [31–33], and we encourage readers to read the work of Fortunato [34] since the number of algorithms to find communities is extremely high.

The community detection algorithm may be classified as *graph partitioning or heuristic algorithms* and *hierarchical clustering*, which in turn may classified as *Divisive clustering* and *Agglomerative clustering.*

The algorithms for graph partitioning consist of dividing the vertexes into k groups of a predefined size. Once fixed the number of elements in each community, it is minimized the number of edges between vertexes of different communities. The number of edges running between clusters is called *cut size.* A classical example of one of these algorithms is the *Kernighan-Lin algorithm* [35].

The other general approach to deal with community detection in graphs is hierarchical clustering. It is based on the identification of groups of vertexes showing a high similarity. Hierarchical clustering algorithms are said to be

divisive if clusters are iteratively split by removing edges connecting vertexes with low similarity, while the *agglomerative* hierarchical algorithms are those which the clusters are merged if their similarity is sufficiently high.

An example of a divisive algorithm is the *Girvan and Newman algorithm* [36], which is an iterative algorithm that removes the edge with the highest betweenness centrality until no edges remain. The result of this algorithm is a *dendogram* with a hierarchical structure of vertexes. In order to obtain the best community partitioning, it is sufficient to calculate a quality measure to each partition and select the one with the highest quality value.

One of the most widespread quality measure for graph partitioning is the *modularity*, that will be explained in Chapter 7.

Finally, an example of an agglomerative hierarchical algorithm is the *Louvain algorithm* [37]. This algorithm starts considering each vertex as a community, and iteratively updates the communities by merging to each community other community which maximizes the modularity of the graph in each iteration (if exists) until a predefined level of modularity is reached.

Blondel's community detection algorithm was used to automatically infer the architecture of a deep learning model in Chapter 7.

3 Journal article (i)

You may delay, but time will not.

Benjamin Franklin.

Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years

Pérez-Benito, F.J.[1,2], Sáez, C.[1], Conejero, J.A.[2], Tortajada, S.[1,3,4], Valdivieso, B.[3], García-Gómez, J.M.[1,3,4]

1 Biomedical Data Science Lab. Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
2 Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
3 Unidad conjunta de investigación en reingeniería de procesos socio-sanitarios. Instituto de Investigación Sanitaria La Fe. Hospital Universitario La Fe, Avenida Fernando Abril Martorell 106, 46026 Valencia, Spain.
4 Red de Investigación en Servicios de Salud en Enfermedades Crónicas (REDISSEC), Valencia, Spain.

Abstract.

Objective: To evaluate the effects of Process-Reengineering interventions on the Electronic Health Records (EHR) of a hospital over 7 years.

Materials and methods: Temporal Variability Assessment (TVA) based on probabilistic data quality assessment was applied to the

historic monthly-batched admission data of Hospital La Fe Valencia, Spain from 2010 to 2016. Routine healthcare data with a complete EHR was expanded by processed variables such as the Charlson Comorbidity Index.

Results: Four Process-Reengineering interventions were detected by quantifiable effects on the EHR: (1) the hospital relocation in 2011 involved progressive reduction of admissions during the next four months, (2) the hospital services re-configuration incremented the number of inter-services transfers, (3) the care-services re-distribution led to transfers between facilities (4) the assignment to the hospital of a new area with 80000 patients in 2015 inspired the discharge to home for follow up and the update of the pre-surgery planned admissions protocol that produced a significant decrease of the patient length of stay.

Discussion: TVA provides an indicator of the effect of process reengineering interventions on healthcare practice. Evaluating the effect of facilities' relocation and increment of citizens (findings 1, 3-4), the impact of strategies (findings 2-3), and gradual changes in protocols (finding 4) may help on the hospital management by optimizing interventions based on their effect on EHRs or on data reuse.

Conclusions: The effects on hospitals EHR due to process reengineering interventions can be evaluated using the TVA methodology. Being aware of conditioned variations in EHR is of the utmost importance for the reliable reuse of routine hospitalization data.

Keywords: Temporal Variability Assessment, Process Reengineering indicator

---

3.1 Background and significance

3.1.1 Introduction

A business process is defined as a structured set of activities performed in any organization for the description of the logical order and dependence of the processes carried out [38]. In healthcare organizations, business Process Reengineering means improving organizational performance by process or information system redesign, covering the needs of healthcare institutions [39–44]. Business process redesign has been applied in many healthcare systems such as pharmacies [45] and emergency departments [46] to increase their efficiency since they are now under pressure all over the world [47]. The authors of the review [48] showed that many of the studies that address the

promotion of business process reengineering in the health sector are related to the reduction in the length of hospitalization or the help with organizational change and how this promotion may drive the development of similar actions, that seek to improve the quality of the services offered, in other organizations.

The data used to evaluate the population's health underlies the effects of the decision-making processes that rely upon these data [49]. When assessing data quality in health systems, one of the most commonly examined dimensions is timeliness [49, 50], which are considered to be an extrinsic data quality concept influencing fitness-to-use features [9, 12].

Our aim was to make a descriptive and retrospective analysis about the process reengineering interventions influence on EHR, and to analyze how these interventions might have influenced hospital activities focusing on the potential technical knowledge which may be extracted from data. The TVA methodology was applied to a database that collects information on admissions to the Hospital Universitario y Politécnico La Fe (HFE) in Valencia between January 2010 and December 2016.

As will be discussed in Section Discussion, many works in recent literature are usually centered in one process and measures how well the intervention is working. Meanwhile, this study count on the main objective of applying a well-documented methodology for the evaluation of temporal variability [4, 5, 51, 52] based on Information Geometry, not only to measure the influence of one process reengineering intervention but also to automatically detect interventions through data distributions.

3.2 Materials and Methods

3.2.1 Ethics

This study did not involve any risk or changes to the healthcare services to patients and did not alter their regular intervention and treatment. Only authorized persons obtained data from electronic health records. They maintained the privacy and security of patients' personal information by encoding their identity with dissociated non-traceable codes. This research was carried out in accordance with the International Guideline for Ethical Review of Epidemiological Studies [53] and the Biomedical Research Ethics Committee of the HFE [54], which approved the study protocol on October 10th, 2017 under the name "ANÁLISIS DE LA CALIDAD Y VARIABILIDAD DE DATOS MÉDICOS" (Registration Number 20170482).

3.2.2 Materials

The study considered the hospitalization data repository of the HFE, in Valencia, Spain, including 108347 admissions from 2010 to 2016. The HFE coordinates all public healthcare services provided by La Fe Valencian Health

Department, from primary to tertiary care, covering 300000 inhabitants directly and adding up to 515000 persons from the catchment area. The HFE is the biggest reference hospital in the Comunitat Valenciana and the fifth largest in Spain. The HFE department is composed by the HFE (with 1000 beds approx.), the health center of Campanar, located in the old facilities of the HFE, the specialty center Ricardo Trènor Palavicino and 20 primary health centers. The health department is met by a team of more than 7000 people, that includes more than 1100 doctors, 400 Internal medicine residents, 3800 positions of different nursing areas and 1,500 people for management and general services.

The repository includes healthcare information on each hospital admission of the overall population during the aforementioned period. After gathering the data, we excluded the episodes of isolated patients, i.e. those who did not belong to the HFE department (for example tourists who are visiting the city), because of the possibility of missing significant information for the study, such as 30-day unplanned readmission or the diagnosis of chronic diseases prior to the date of admission.

Before conducting the TVA, a preprocess was carried out on some administrative and clinical variables. The original dataset was completed with some aggregate and processed variables, including the age of the patient which was computed as the difference between the admission date and their birth date, the Charlson comorbidity index score [55] that was calculated using updated weights from Schneeweiss, [56] and the ICD-9-CM coding, as proposed by Quan, [57]. This score was calculated by adding 1 point for the patient's history of acute myocardial infarction, peripheral vascular disease, cerebrovascular disease and diabetes without complications; 2 points for congestive heart failure, chronic obstructive pulmonary disease, mild liver disease, diabetes with complications and malignancy; 3 points for dementia and renal disease; 4 points for moderate-to-severe liver disease and HIV infection; and 6 points for metastatic cancer. This score was calculated using another repository that included the ICD-9-CM code for each diagnosis of chronic disease recorded in the HFE.

The list of variables considered is shown in Table 3.1. Extra information on the materials can be found in the S1 Appendix.

Variables	Description	Type(values/format)
Sex	Sex of the person	Discrete (Male, Female)
Age	Age in years at the time of the admission	Numerical Integer
AdmissionServiceCode	Code of the service of hospitalization	Discrete 4-length alphanumeric code
RealServiceCode	Code of the service related to the episode	Discrete 4-length alphanumeric code
DischargeServiceCode	Code of the service which discharged the patient	Discrete 4-length alphanumeric code
AdmissionReason	Reason for hospital admission	Discrete (See S1 Appendix)
DischargeDate	Date of patient discharge	Date (yyyy/mm/dd)
DischargeTurn[1,2,3]	Discharge shift	Discrete
DischargeReason	Reason for patient discharge	Discrete (See S1 Appendix)
DischargeDestination	Destination after patient discharge	Discrete (See S1 Appendix)
DischargeBefore12	Discharge before 12:00 noon	Discrete (Yes, No)
Exitus	Death of the patient during hospitalization	Discrete (Yes, No)
Exitus48	Death of the patient within two days after hospitalization	Discrete (Yes, No)
HospitalTransfer	Existence of hospital transfer	Discrete (See S1 Appendix)
LengthOfStay	Length of stay of hospitalization episode. It is measured by the number of nights that the patient was admitted	Numerical Integer
Intervention	Surgical Intervention	Discrete (Yes, No)
PreoperatoryStay	Length of stay before the intervention	Numerical Integer
Readmission30	Was the patient readmitted during the 30 days after discharge?	Discrete (Yes, No)
CharlsonIndex	Charlson comorbidity index for hospitalization	Numerical Integer

Table 3.1: **List of variables contained in the study case.** The shift in which the patient is admitted and discharged is coded as 1 for the morning (from 8:00 am to 3:59 pm), 2 for the evening (from 4:00 pm to 11:59 pm) and 3 for the night (from 0:00 to 7:59 am)

3.2.3 Methods

3.2.3.1 Theoretical Background

A systematic TVA methodology based on probabilistic data quality control was applied [5, 18, 58]. This methodology uses methods based on Information Geometry [16, 59] which provide a way for the comparison of dissimilarities between probability distributions of different temporal data batches.

It firstly consists of modeling Probability Density Functions (PDF's) -in our case, it was made by the use of Kernel Density Estimation [60]-. The Jensen-Shannon distance (JSD), which is a symmetrized and smoothed version of the Kullback-Leibler divergence [61, 62], provides a way to measure how different the non-parametric PDF's are.

The space in which each point represents one PDF and the distance between two points is that defined by the aforementioned distance, forms a simplex and is known as statistical manifold and possesses good mathematical properties [16].

This function representation allows us, for example, to compute the centroid of the PDF's and to apply projection methods, such as Principal Component analysis [63] or Multidimensional Scaling [64, 65]. These artifacts, as can be seen in Figure 3.1 where a short artificial experiment was driven

to yield a simple proof of concept, give us the possibility of quantifying the dispersion and making space representations as a graphical way to detect variability. The exploratory methods provided by the methodology are:

- **Information Geometry Temporal (IGT) plot:** This presents a visualization of the temporal evolution of data. Temporal batches are laid out as a 2D plot while conserving the dissimilarities among their distributions. The IGT plot helps to reveal temporal trends in the data (as a continuous flow of points), abrupt changes (as an abrupt break in the flow of points), recurrent changes (as a recursive flow through specific areas), conceptually related time periods (as grouped points) and punctual anomalies (as isolated outlying points). Temporal batches are also labeled to show their date. They give seasonal information by means of colored labels (warm colors for summer and cool colors for winter) and are supported by a smoothed timeline path joining them [18]. The Density-based spatial clustering of applications with noise (DBScan) [66], was applied to the IGT plot using the median of the JSDs as grouping coefficient in order to automatically find temporal groups.
- **Probabilistic Statistical Process Control (PDF-SPC) algorithm:** The purpose of PDF-SPC is to monitor the degree of change in data variability distributions throughout consecutive temporal batches (in our case months), to a moving reference distribution -initially the first batch. According to the magnitude of the current change, measured by the JSD with respect to the reference distribution, the degree of change of the repository is classified into three states: in-control (distributions are stable), warning (distributions are changing), and out-of-control (recent distributions are significantly dissimilar to the reference, leading to an unstable state and yielding a change in the reference distribution). When an out-of-control state is reached, a significant change is confirmed and the reference distribution is set to the current one for subsequent comparisons. The warning and out-of-control states are represented as broken and continuous vertical lines, respectively.
- **Temporal Heat Maps:** Temporal Heat maps show the absolute or relative frequencies over time. The Temporal Heat map of a variable is a 2D plot in which the X-axis represents the time, the Y-axis represents a possible data value or range of values of the variable, and the color of the pixel at a given (X, Y) position indicates the frequency at which value Y was observed on date X. These heat maps facilitate a rapid broad visualization of the evolution over time of the values of the given variable.

The TVA methodology consists of using these methods iteratively. In a top-down approach, we start by analyzing the temporal variability of the complete monthly-batched data set. We then drill down to the specific variables or groups of variables which best explain the variability detected, according to the results of the analysis and prior knowledge of the repository.

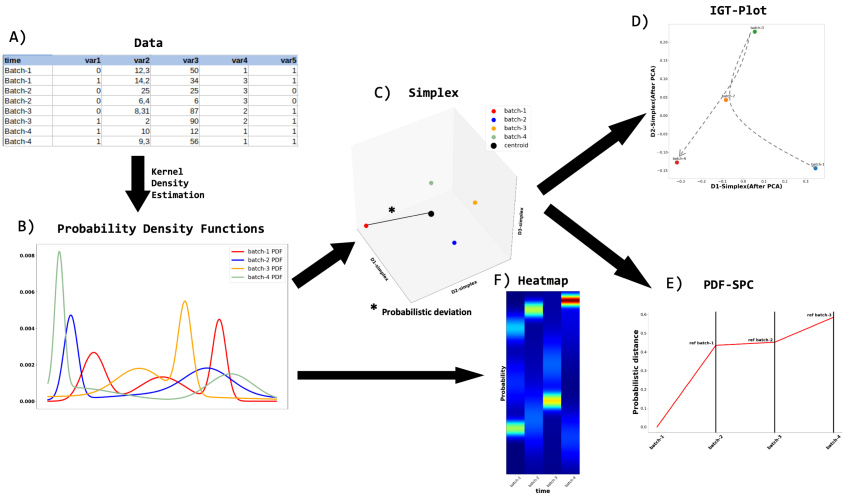


Fig. 3.1: **Technical diagram.** The TVA methodology is based on information geometry. A short artificial experiment taking 4 temporal batches (only 4 batches were taken to ensure that the simplex could be represented in three dimensions) was drawn with the purpose of clarifying the concept. A) represents the generated artificial database in which binary, quantitative and continuous variables cope with. B) is the PDF representation. C) shows the simplex in which each point represents the PDF of one batch and the bigger black point represent the centroid of the simplex (the distance from each batch to the centroid serves as a dispersion measure). D) is the IGT-plot, in studies with more batches is one way to graphically represent the variability among the batches and to apply clustering methods to automatically detect temporal patterns, it must be noted that the color changes from previous representations to simulate the seasonal color mapping. E) shows the PDF-SPC, since the database was designed to present high variability, all the batches are "out-of-control". Finally, F) presents the heatmap of the concatenated batches distributions which allows monitoring temporal pattern changes.

3.2.3.2 Working methodology

This study was carried out by a multidisciplinary team of professionals from various fields: the technical background was provided by a computer scientist, a statistician, a mathematician, and specialist physicians whose expertise is the PR and the management of the hospital.

The study protocol was divided into two stages: in the first changes were detected and in the second one, they were analyzed and their causes were searched.

An overview of the study protocol is shown in Fig 2, in which the iterative protocol used for the detection of process reengineering interventions is described.

Following the previously described TVA methodology, we start by considering the whole multivariate dataset grouped by monthly batches under the assumption that PR interventions may imply an impact on EHR. This is intended to detect different data behavior patterns (see A) in Figure 3.2). Secondly, the same methodology was applied to the detected temporal data changes with a univariate approach to identify the variable, or set of variables, that could have influenced the observed global change. Subsequent automatic iterations for each variable may identify more univariate pattern changes which could have been smoothed due to multivariate batches with a greater global impact. These iterations can also detect interactions in the variables produced by changes in one variable (see B) in Figure 3.2).

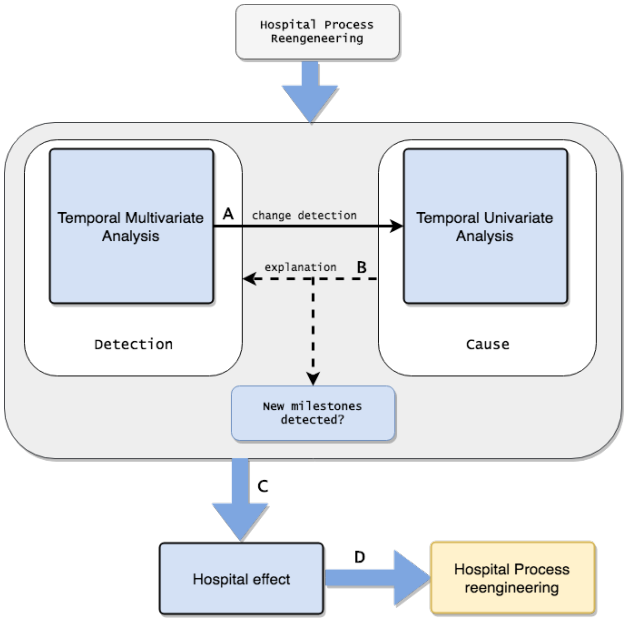


Fig. 3.2: **Work-flow diagram.** Multivariate analysis is able to discover changes driven by the global probabilistic variability A). The obtained findings drive us to make the univariate analysis with the purpose of explaining the aforementioned changes B). It is worth mentioning that this step detects smoothed changes which had been covered by more abrupt global differences. Step C) is the evaluation of the interventions which provoked the data change and their implications. Finally, this evaluation could serve as the starting point for the implementation of PR D).

3.3 Results

We provide the description of how the proposed methodology was able to detect the effects, through data analysis, of the process reengineering interventions which will be shown in Section 3.4.1. The list of findings mapped to the PR interventions carried out in the hospital in that period is shown in Table 3.2 in which the numerical evidence was added.

Finding	Intervention	Evidence
F1	I1 - Hospital relocation	■ IGT-Plot and its DBScan clustering show differences between 2010 and the rest of years of the study (Figure 3.3 (Multivariate $JSD(10D, 11Jan) = 0.74$, 2010 belongs to the green cluster and the rest of years belong to the blue cluster) ■ Heat map of the PCA dimension reduction of the multivariate analysis offers absolutely different pattern during the months of 2010 (Figure 3.3) ■ PDF-SPC of AdmissionService, DischargeService, and RealService show the abrupt change detected at the end of 2010 (Figure 3.4). (AdmissionService univariate $JSD(10Jan, 11Jan) = 0.26$, DischargeService univariate $JSD(10Jan, 11Jan) = 0.26$ and RealService univariate $JSD(10Jan, 11Jan) = 0.25$)
F2	I2 - Services reconfiguration	■ The heat map shows a trend of refinement of the red central band in the closest months to February 2011 (Figure 3.3). ■ February 2011 is detected as an outlier by the IGT-plot and its DBScan clustering (Figure 3.3) (Multivariate $JSD(11Jan, 11F) = 0.41$ and Multivariate $JSD(11F, 11m) = 0.73$, besides DBScan did not assign to any cluster) ■ PDF-SPC of AdmissionService, DischargeService, and RealService show the abrupt change detected at the end of 2010 (Figure 3.4). (AdmissionService univariate $JSD(11Jan, 11m) = 0.33$, DischargeService univariate $JSD(11Jan, 11m) = 0.29$ and RealService univariate $JSD(11Jan, 11m) = 0.29$)
F3 [1]	I3 - Care services reconfiguration	■ Heat map marks a different pattern of three months in mid-2013 (Figure 3.3). ■ PDF-SPC of AdmissionService, DischargeService and RealService show the abrupt change detected in mid-2013 (Figure 3.4). (AdmissionService univariate $JSD(13M, 11m) = 0.34$, DischargeService univariate $JSD(13M, 11m) = 0.34$ and RealService univariate $JSD(13M, 11m) = 0.34$) ■ PDF-SPC, IGT-Plot and its DBScan clustering display an abrupt change in mid-2013 (Figure 3.5). (DischargeDestination univariate $JSD(11a, 13m) = 0.29$, two clusters well-defined of months prior to March 2013 and the rest.
F4	I4 - Inclusion of 80000 patients. The update of the pre-surgery admission protocol	■ DBScan applied to IGT-plot warns of the existence of a month -January 2014- with an atypical behavior (Figure 3.3) (Unassigned month). ■ This atypical month is also detected by the Heat map (Figure 3.3). ■ PDF-SPC, IGT-Plot and its DBScan clustering show that a change occurred in early 2014 (Figure 3.6) (DischargeDestination univariate $JSD(11m, 14F) = 0.22$, two clusters well-defined of months prior to February 2014 and the rest). ■ January 2015 is also detected as an atypical month by IGT-plot and its DBScan clustering (Figure 3.3). (Unassigned month). ■ The heat map analysis reveals that the number of hospitalizations increased from this month. It can be seen thanks to the width of the red band and is supported by the increment of admissions detected in 2015 (Table 3.3).
F5[1]	I3	■ PDF-SPC, IGT-Plot and its DBScan clustering show an abrupt change in mid-2016 (Figure 3.5) (DischargeDestination univariate $JSD(14F, 16M) = 0.24$)

Table 3.2: These findings were directly observed from data after the application of the methodology described in Section 3.2.3.
[1] The finding F5 was a direct cause of the intervention carried out and detected by the finding F3 (it will be discussed after).

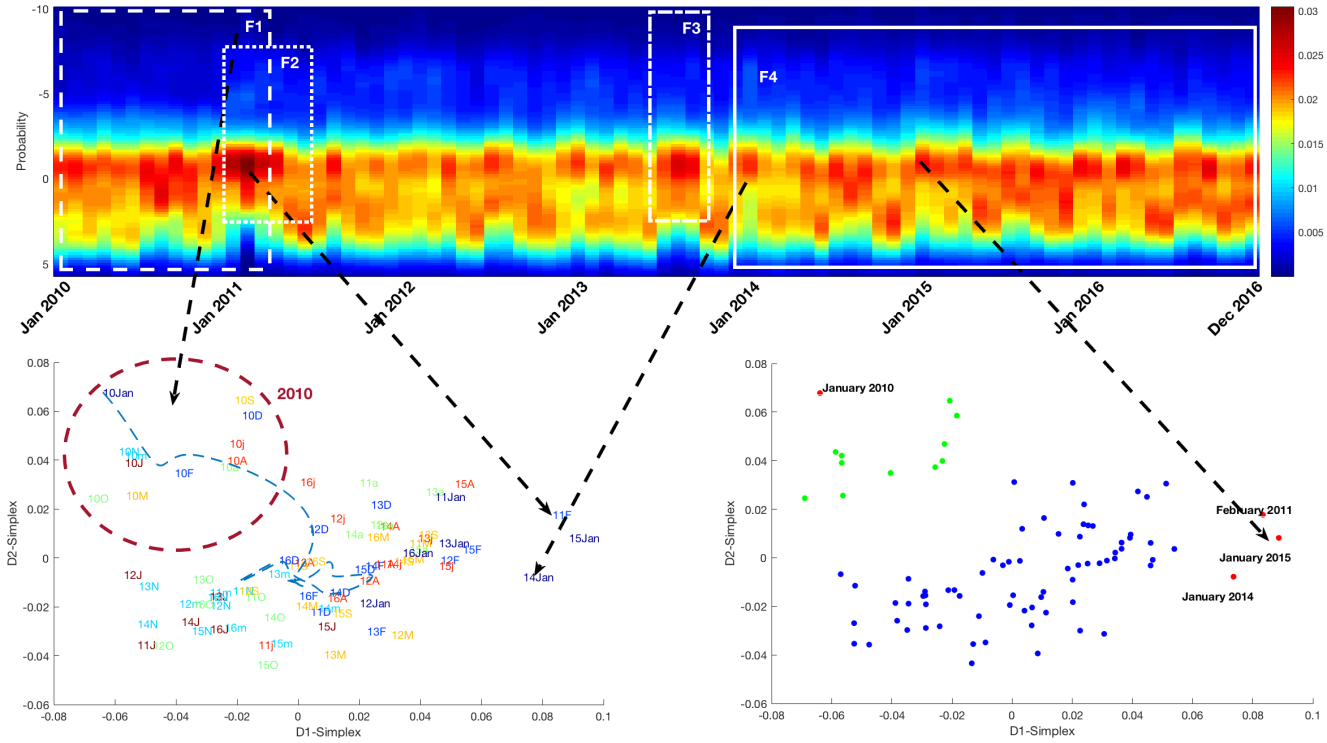


Fig. 3.3: **Multivariate analysis of hospitalizations in HFE.** Four findings were detected. The top figure presents the heat map of the temporal multivariate data distribution, down left figure shows the IGT plot where the whole set of variables were considered and finally the DBScan clustering of IGT plot is exhibited in right down figure. F1 is correlated with the difference between 2010 and the rest of the years; F2 is aligned to data changes in early 2011; F3 stands for three months in mid 2013 with atypical patterns; F4 refers to January 2014, which is quite different from other months and introduce the beginning of an atypical pattern; the outlier detected in January 2015 is the precursor of the increase of frequency observed in the subsequent months. By analyzing the IGT-Plot and its clustering, we discovered that the heat map of the one dimensional PCA presented temporal color patterns.

In Figure 3.3 it can also be seen that the HFE probably suffered at least one important change in late 2010 (F1) and early 2011 (F2) that caused an abrupt change in all the monthly variable distributions, another significant event can be detected at mid-2013 (F3) where a density condensation may be observed on the top of the maximum-frequency band. Finally, at the beginning of 2014 (F4), an atypical month is detected, this month is followed by temporal patterns that had not been observed before and that led to an increment of the frequency from early 2015, 2015 January is detected as an outlier in the right down picture in Figure 3.3.

The univariate methodology was used in pursuit of an explanation for these changes. The changes can be explained by almost all the variables.

Figure 3.4 shows that the variables which store the admission, real and discharge service of each hospital episode explain F1 and F2. The configuration of the hospital services may also explain Finding F3. The PDF-SPC's of the services configuration is shown in this figure.

After removing the cases prior to March 2011, the same methodology was applied in order to avoid the non-detection of findings by the smoothing, which could have caused the abrupt changes prior to this date. As already mentioned, the changes previous to March 2011 had an impact on the whole set of variables.

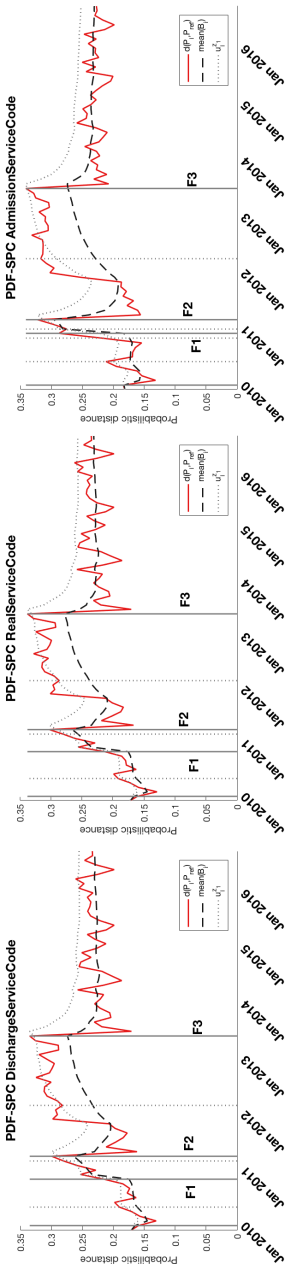


Fig. 3.4: **PDF-SPC of the three variables related to services configuration (Admission, Real and Discharge Service).** Findings F1, F2, and F3 are detected in the three variables by out-of-control states.

Figure 3.5 shows the PDF-SPC, IGT Plot and its DBscan clustering for the variable DischargeDestination, showing the change of the discharge policy introduced between early-mid 2013 and mid-2016 which will be discussed in the next Section. This change is probably related to Finding F3 and will be referred to as Finding F5 (F5 is a new milestone -understanding milestone as different data distribution pattern- detected by applying the univariate methodology. The emergence of new milestones can be seen in Figure 3.2. Two temporal clusters were found.

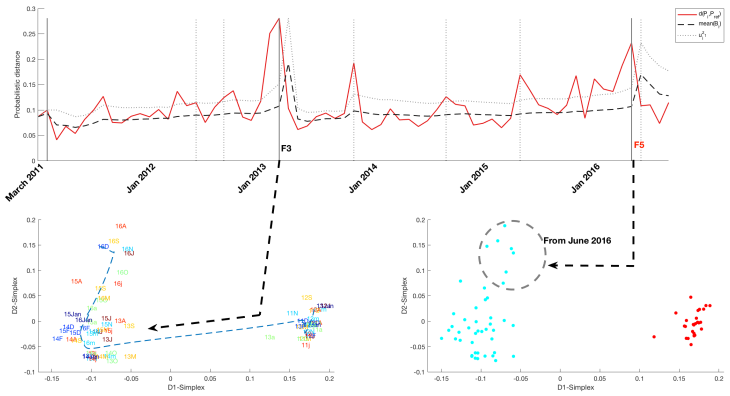


Fig. 3.5: **PDF-SPC (top figure), IGT plot (left down figure) and its clustering by DBScan (right down figure) of the variable which records the method of patient follow-up after discharge (DischargeDestination).** The analysis of this variable shows new evidence for Finding F3 as well as a new Finding F5 (a new milestone is detected as mentioned in Figure 3.1 which probably was not detected by the multivariate analysis due to the higher hospitalizations from 2015 January.

The exploratory PDF-SPC visualizations, IGT Plot and its DBscan clustering for the LengthOfStay variable are shown in Figure 3.6, where a variation in the patients' average length of stay in early 2014 can be seen correlated with Finding F4. The histograms of this variable show an increase of 1-day stays with respect to 2 and 3-day stays (see S7 Figure).

The number of annual hospital admissions is shown in Table 3.3. It can be seen that the number of patients increased significantly from 2015 and this could have caused the change detected in the multivariate analysis (see Figure 3.3).

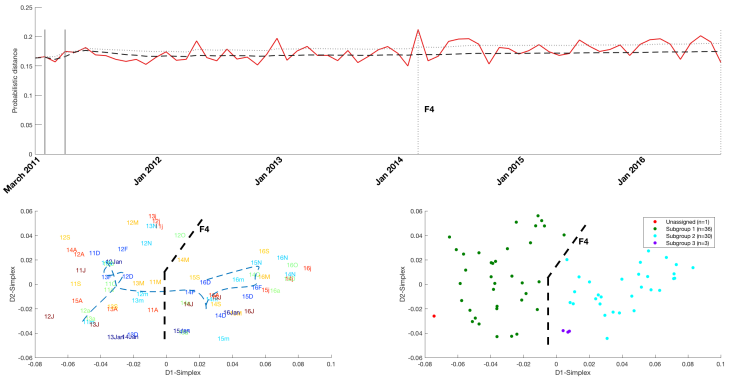


Fig. 3.6: **PDF-SPC (top figure), IGT plot (left down figure) and its clustering by DBScan (right down figure) of the variable which measures the number of hospitalization days (LengthOfStay).** A change in the length of stay occurred in early 2014, related to Finding F4 was discovered in the multivariate analysis.

Year	Number of admissions
2010	14706
2011	12969
2012	14212
2013	14459
2014	14259
2015	18063
2016	19643
Total	108347

Table 3.3: **Hospital admissions flow.** Number of patients admitted per year

3.4 Discussion

Healthcare organizations are constantly forced to increase the quality of care while maintaining an optimum use of resources [39, 42]. Therefore, managerial decisions, which are routinely taken in a business environment, are constantly influencing data distributions. These decisions may imply temporal variability inherent to the data. In this field, the impact may not only be on the hospital management, but also on the regular population health and on the perception of its quality [67, 68].

There exist some approaches to carry out the assessment of process reengineering interventions on literature. The authors of [69] propose a methodology

based on process-mining to measure the organizational changes in the stroke emergency process. The assessment was performed by the use of PALIA [70]. One of the most powerful tools for process-mining is PROM [71] which covers a wide range of process-mining algorithms such as α-algorithm [72], genetic process-mining [73] or Heuristic Miner [74].

The present study searches something similar but is based on a consistent methodology driven by the variations of PDF applied to a health service dataset with the purpose of studying the effect of PR interventions on data. This methodology is used to monitor data distributions through time, becoming a way for "real-time" detection of the impact of management decisions and process reengineering interventions on hospital activities as well as finding undesired factors or effects [75]. We think that the principal Impact of our methodology is its global applicability when compared to the aforementioned approaches. These approaches are usually centered on one process and measures how well the intervention is working. Meanwhile, the proposed methodology provides the capability of detecting the own interventions by the multivariate iteration and its influence (not only direct but also indirect) in other related processes by the univariate iteration.

Besides, another contribution is that the detection of data distribution changes can lead to the improvement of future decisions and research work, for instance, a 30-day readmission model or the development of longitudinal studies could be better built from the prior knowledge of the findings of our study.

Although some of the milestones that have been detected are not the result of process reengineering, but rather are specific daily situations that influence the operation of the hospital. These milestones have been taught because we believe that these situations could motivate one or more interventions in terms of process reengineering. We also remark that HFE experts in PR analyze the impact on hospital management as well as on the regular patient population's health by exploring the reasons and the effects on hospital activities of decisions already taken (see C) in Figure 3.2). Although the following is outside the scope of this work, the results of this analysis may help to identify indicators which could be the input for further PR decisions (see D) in Figure 3.2).

A list of the process reengineering interventions, contributions, limitations and lines of future work are given below.

3.4.1 Process reengineering interventions

The process reengineering interventions carried out by the hospital managers and their motivation are presented in chronological order with the purpose of correctly interpreting the findings, shown in Section Results, that the exploratory method applied was able to detect.

3.4.1.1 Hospital relocation (I1)

The HFE relocated to new facilities between December 2010 and February 2011, which involved a progressive reduction of admissions that lasted while the intervention finished, the time when hospital activity recovered. The relocation protocol was the following:

- The Outpatient Department was relocated on November 2010. The first allergy, dermatology, internal medicine, and infectious disease consultations took place on November 29th (finding F1).

The remaining areas were progressively moved from lower to higher logistical complexity. Finishing with the transfer of the most delicate areas as follow:

- Maternity and Child Health was transferred on February 13th, 2011, moving 81 children and premature babies and 11 pregnant women and recently delivered mothers (finding F2).
- The adult hospitalizations area was relocated on February 20th, 2011, with 158 adults (finding F2).

Consequently, after December the admission-patient typology became urgent profiles (see S1 and S2 Figures). The admission of patients with a higher age and comorbidity index was caused by the relocation since this type of patient frequently has a serious illness and requires more urgent resources. The number of interventions decreased, as allowed for in the managers' planning (see S2 Figure). After opening the new facilities, more hospital transfers were (see S3 Figure) needed and the information system was changed. The admission planning taken by the hospital management for the relocation was quite similar to the interventions adopted during the summer months, in order to allow for staff holidays, which can also be detected by the seasonality in the data.

3.4.1.2 Services re-configuration (I2)

At the beginning of 2011 (Finding F2) and closely related to the previous point, the services were restructured (see S4 Fig) when the old facilities composed of four hospital centers were combined into one. The services were reorganized into clinical management areas and a committee for the approval or rejection of changes in service configuration was created.

3.4.1.3 Care services distribution (I3)

Despite the abovementioned relocation, some of the patients were still treated in the old facilities, as in the case of chronic patients, since it was decided to send them to the old facilities for patient follow-up at the beginning of 2013. This intervention involved a new service re-structuring and a higher

percentage of patients were sent to their general practitioner in detriment of those discharged home for follow-up (finding F3, see Table 3.2). This situation was temporary due to the closure of the previous chronic unit in mid-2016 (finding F5, see Table 3.2), which meant more patients were monitored at home. A higher quantity of resources, therefore, had to be allocated to this end (see S5 Figure). Figire 3.3 served to detect this intervention and allowed us to suspect that a new cluster would probably have appeared from mid-2016 if the following months had been added to the dataset.

3.4.1.4 Changes in the pre-surgery admission protocol due to the inclusion of patients from another hospital (I4)

Another important intervention adopted by the hospital management in 2015 was due to prior knowledge of the assignment to the HFE of approximately 80000 patients (now the hospital covers around 280000 inhabitants when before were approximately 200000 citizens) previously assigned to another Valencian Hospital, the Hospital Doctor Peset (finding F4). The uptake of this population was expected to initially cause an increase of 50 daily urgent admissions, progressively rising to 70. For this reason, three actions were taken, in which we can also find some of the findings previously detected by the methodology:

- A new surgical admissions unit was created to assess patients to be hospitalized.
- The number of beds assigned to home hospitalization was increased (see S6 Figure) to cover two more areas (Pediatrics and Neonatology) (previously chronic, mental health and pediatric oncology patients).

The pre-surgery planned-admission protocol was updated in early 2014. Whereas before this intervention, patients were admitted the night previous to surgery, they were now admitted on the morning of the intervention and they had a bed ready at midday after daily patient discharges. This meant an increase of the daily bed-occupation in the hospital and also on patient satisfaction, due to the shortening of the stay. The isolation of January 2014 in the multivariate analysis (see Figure 3.3) was probably caused by this change.

3.4.2 Discoveries and possible particular contributions

Time is a factor which has been studied as part of data quality dimension, generally leading to dimensions such as timeliness, currency, volatility, concordance or comparability [76–80]. Some of the data quality dimensions are used for validation of the quality of care [81]. The general contribution obtained by the TVA proposed is the use of the assessment of a data quality dimension in the monitoring of the interventions carried out by the hospital. For each intervention we want to highlight:

- **I1.** The relocation of the hospital. More than 1800 professionals were involved in the operation and 40 ambulances were needed for the transfer. The data suffered a great impact, both multivariate and univariate (see Section 3.3 and S1, S3 Figures) an the impact on the whole set of variables was monitored by the TVA proposed in the present study. The impact on the variables was produced not only in the expected ones. In this sense, the TVA monitoring may provide an added value when is used as a tool for "real-time" detection.
- **I2.** Services reconfiguration due to changes in hospital management policies by logistic relocations (see Section 3.4.1.2). Some changes in services and treatment areas occurred during the study period. In addition to its capacity for management process control, our proposed methodology can reveal information and subsequent considerations to help in data reuse, for example for prediction purposes as well as for observational studies involving the comparison of different services during a period of reduced data quality.
- **I3.** Reconfiguration of care areas due to PR decisions (See Section 3.4.1.3). After a logistic relocation, the hospital activity probably suffered several unexpected difficulties. These difficulties led to PR decision-making that can be monitored by the proposed TVA which may be useful to create "PR effectiveness indicators" to be used as a background for future interventions.
- **I4.** The inclusion of 80000 patients from another Valencian Hospital. It would have produced a hospital overcrowding if the interventions (detected by the proposed approach) had not been taken. The most important intervention produced an increase in the percentage of surgeries carried out on the day of admission, rising from 0% to 75%, avoiding a collapse due to an increase in the percentage of beds occupied, which rose to 97% from the previous 82%. It is worth mentioning that one of the challenges in the rise in the number of patients was the integration of computer data into the Business Intelligence used by the hospital. The knowledge of both the increased population assigned to the HFE and the pre-surgery planned-admission protocol change may influence the corresponding data for descriptive or research purposes.

3.4.3 Limitations

One of the principal advantages of the TVA methodology used here is its capacity to analyze a great number of variables in a single iteration. This may also influence the loss of information about what is happening and where at a higher granularity, implying the need for knowledge in the field of study. For instance, finding 4 presented in Section Results firstly was considered as two findings, the hospital PR expertise was needed to understand the scope of the intervention associated with this finding.

Using a single-component PCA reduces the dimensionality of the iteration of the multivariate analysis and may smooth other discoveries with less impact in global terms, making the univariate iteration necessary not only to explain but also to detect. The use of other non-linear reduction methods such as t-distributed Stochastic Neighbor Embedding [82] or machine learning approaches [83, 84] may have a better fit in certain cases and also contribute in the pursuit of interactions between variables.

Faulty healthcare processes are one of the main causes of practitioners making technical mistakes [85], can compromise patient safety and even cost lives [86]. However, in this study, we did not focus our attention on detecting processes for improvement, which is another possible application of the methodology for healthcare management.

3.4.4 Future work

In line with the present study, and to overcome the limitation mentioned above, we aim to develop an automated algorithm that can suggest the origin of the multivariate changes in terms of a set of implicated variables or their interactions.

3.5 Conclusions

Temporal variability in EHR may be considered as an intrinsic data quality feature due to its implications for data reuse. In this work, we have demonstrated how data changes over time and how the statistical distributions of EHR are biased by clinical and management PR interventions in the case of a Valencian hospital over seven years. Analyzing the temporal data variability by means of TVA has the potential not only to detect but also to monitor Big Data hospitalization resources, in order to improve the assessment of PR in healthcare systems.

Acronyms for months

Jan: January, F: February, m: March, a: April, M: May, j: June, J: July, A: August, S: September, O: October, N: November, D: December.

3.6 Supplementary material

S1 Appendix

This document presents the consort diagram for the data base considered for this study, the tables needed for the comprehension of the multi-category variables -understanding multi-category variables with more than three options-.

Data base description

1. Consort Diagram

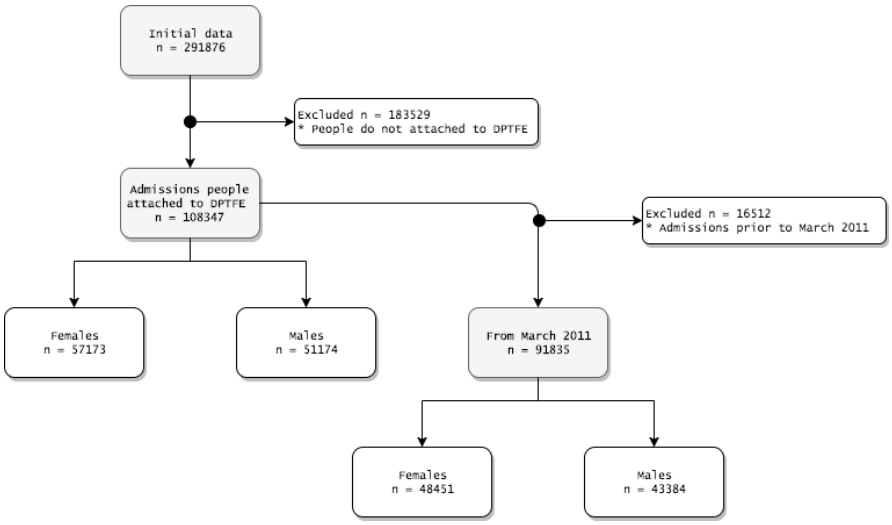


Fig. A1: Consolidated Standard Reporting Trials (CONSORT) flow diagram of the case study of the HFE hospitalization data base.

2. Dictionaries

Service codes. Used for the variables *AdmissionServiceCode*, *RealServiceCode* and *DischargeServiceCode*:

Code	Description	Code	Description	Code	Description
HUDO	Pain unit	HURO	Urology	HREI	Child rheumatology
HPSA	Adolescent psychiatry	HHEM	Hematology and hemotherapy	HCGI	Pediatrics surgery
HREU	Rheumatology	HCCV	Cardiovascular surgery	HUTA	Eating disorders unit
HORP	Orthoptics and pleoptica	HTRA	Traumatology	HRTE	Radiation oncology
HDII	Child Digestive	HNAI	Child pneumology and allergies	HCMX	Maxillofacial surgery
HMUR	Short stay unit	HCG2	General surgery and digestive system II	HSEP	Septic unit
HECI	Child endocrinology	HNEN	Neonatology	HCDG	Digestive surgery
HNCI	Child neurosurgery	HCUR	Curietherapy	HCPI	Child plastic surgery
HCVI	Child cardiovascular surgery	HUHP	Hepatobiliopancreatic	HPED	General pediatrics
HRXD	Radiology	HMIN	Internal medicine	HALI	Child allergy
HRQU	Burn care resuscitation	HOFT	Ophtalmology	HCAR	Cardiology
HGIN	Gynecology	HONI	Child oncology	HQUE	Burn care
HUEG	Esophagogastric surgery	HCTO	Thoracic surgery	HSII	Child psychiatry
HREP	Reproduction	HCIR	General surgery and digestive system	HMNU	Nuclear medicine
HRHB	Rehabilitation	HUHT	Hemostasis and thrombosis	HCAI	Child cardiology
HECR	Endogrinology and nutrition	HNCG	Neurosurgery	HCOT	Orthopedic and traumatology
HEPR	Refractory epilepsy	HOFI	Child ophtalmolgy	HMDH	Hepatology
HUCP	Pediatric ICU	HRER	Anesthesia-resuscitation (RET)	HUEI	Infectious diseases unit
HORL	Otolaryngology	HUEM	Metabolic endocrine surgery	HUMI	Intensive medicine
HPSI	Psyquiatry	HOBS	Obstetrics	HCLP	Coloproctology surgery
HUMM	Functional Breast Cancer Unit	HNER	Neurology	HALE	Allergy
HPIN	Pediatric infectious	HURQ	Raquis unit	HHMI	Child hematology
HCVA	Angiology and vascular surgery	HDER	Dermatology	HNMI	Child pneumology
HURI	Child urology	HOTI	Child orthopedic and traumatology	HCMI	Child maxillofacial surgery
HMDG	Gastroenterology	HCLP	Plastic surgery	HCEP	Short stay unit and wall
HORI	Child otolaryngology	HUML	Medium-long stay unit	HNEM	Pneumology
HONC	Oncology	HNRI	Neuropediatrics	HULM	Spinal cord injury unit
HMIF	Lower limb unit	HREM	Anesthesia-resuscitation (MAT)	HUTP	Lung transplantation unit
HMET	Child metabolic diseases	HNEF	Child rheumatology	HLIT	Lithotripsy
HNFI	Child nephrology	HREA	Resuscitation	UCSI	Non-admittance surgical unit

Table 1: List of variables contained in the study. Service Identifier

Hospital Codes. Used for variable *HospitalTransfer*:

Code	Description	Code	Description	Code	Description
1000	Inst. Oftalmologico Valencia	1101	Hospital de Vinaros	1102	Hospital Gral. de Castellón
1103	Hospital de la Magdalena	1104	Hospital de la Plana	1111	Hospital 9 de Octubre
1150	Hospital Prov. de Castellón	1191	Termalismo Heliomar Benicassim	1192	Mutua de Azulejeros, Onda
1369	Hospital Intern. Medimar	1537	Hospital Rey Don Jaime	1717	Hospital S. Jaime Torrevieja
1901	Clínica Santa Teresa	1902	Ntra. Sra. de la Misericordia	1903	Ctro de Rehabilitación, Onda
2021	Hospital de Levante	2101	Hospital de Sagunto	2102	Hospital Arnau Vilanova
2103	Hospital Doctor Moliner	2104	Hospital de Requena	2105	Hospital La Fe (Campanar)
2106	Hospital Dr Peset I Aleixandre	2107	Hospital Clínico Universitario	2108	Hospital Malvarrosa
2109	Hospital Francesc de Borja	2110	Hospital Lluis Alcanys Xativa	2111	Hospital d'Ontinyent
2115	Hospital Rehabilitación La Fe	2125	Hospital Maternal La Fe	2135	Hospital Infantil La Fe
2145	Hospital La Fe. Esc Enfermería	2150	Hospital General de Valencia	2155	La Fe Bulevar
2169	Vissum Inst. Oftalmológico	2191	Centro Rehabilitación Levante	2192	Inst. Valenciano Oncología
2194	Hospital Santa Lucía	2526	Hospital de Torrevieja	2837	Hospital de Manises
2840	Hospital de Denia	2844	Hospital la Pedrera	2901	Clínica Blanquer
2902	Cl. Sagrada Familia	2903	Casa de Reposo San Onofre	2904	Clínica Virgen del Consuelo
2905	Hospital de Valencia al mar	2906	Clinica Quirón de Valencia	2907	Clínica Casa de la Salud
2908	Hospital de Mislata (Militar)	2909	Hospital Santa Lucía	2910	Hospital Padre Jofre
2993	FISABIO	3050	Hospital del Vinalopo	3082	Hospital de Lliria
3101	Hospital Marina Alta	3102	Hospital Vila-Joiosa	3103	Hospital Verge del Lliris
3104	Hospital de Elda	1105	Hospital General de Alicante	3106	Hospital S. Vicent del Raspeig
3107	Hospital D'Elx	3108	Hospital de Orihuela	3110	Hospital Sant Joan. Alicante
3116	Fundacion Lluis Alcanyis	3150	Hospital Provincial de Alicante	3191	S.S. FCO. de Borja. Fontiles
3192	Clinica Vistahermosa	3193	Clínica Velazquez II. Alicante	3194	S. Perpetuo Socorro. Alicante
3901	Centro Médico San Carlos	3902	Sanatorio San Jorge	3903	Clínica Benidorm
3904	Instituto Geriátrico de Levant	3905	Policlínico S. Carlos de Denia	3906	Clínica Oftalmológica Buigues
3907	Cl. Médico Quirúrgica C. Jardín	3908	Cl. Villamartin de Orihuela	3909	Psiq. Penitenciario Fontcalent
3910	Hospital PSIQ. and Provincial Alicant	3911	Levante Mediterranea Matepss	7777	Hospital de la Ribera
7778	Hospital 9 de Octubre	7779	Other Country	8888	F.Oftalmológica Mediterránea
9040	Hospital La Fe Unificado	9998	Barraca de Aguas Vivas	9999	Other Community

Table 2: List of hospitals contained in the HFE registries.

Admissions reasons. This variable shows the reason for hospitalization:

Code	Description
0	Undetermined
1	Examination
2	Common disease
4	Accident at work
5	Casual accident
6	Self-injury
7	Aggression
8	Birth
9	Others
10	Pathological new-born
11	Urgent outpatient complication
12	Surgery complication
13	Day hospital complication
14	Interventionism complication
15	Infract
19	Urgent outpatient complication
20	Patient from other hospital planned
60	An influenza research case
61	An influenza probable case
62	An influenza confirmed case
63	An influenza dismissed case
90	UCSI episode complication
99	Disaster

Table 3: List of hospitalization reasons contained in the study.

Discharge reasons. The information collected in this variable indicates the reason for discharge:

Code	Description
1	Healing or improvement
2	Voluntary discharge
3	Transfer
4	Exitus
5	Other
6	In extremis

Table 4: List of discharge reasons contained in the study.

Destination after discharge. This variable contains the patient follow-up method after discharge:

Code	Description
1	Day hospital
2	Discharge home
3	Outpatient care
5	Specialty center
6	Emergency department
8	Escaped
9	Others
10	Medium and long stay hospitals
11	Nursing home or socio-health center
12	General practitioner
13	Disciplinary discharge

Table 5: List of discharge destinations contained in the study.

S1 Figure. Supplementary evidence for I1.

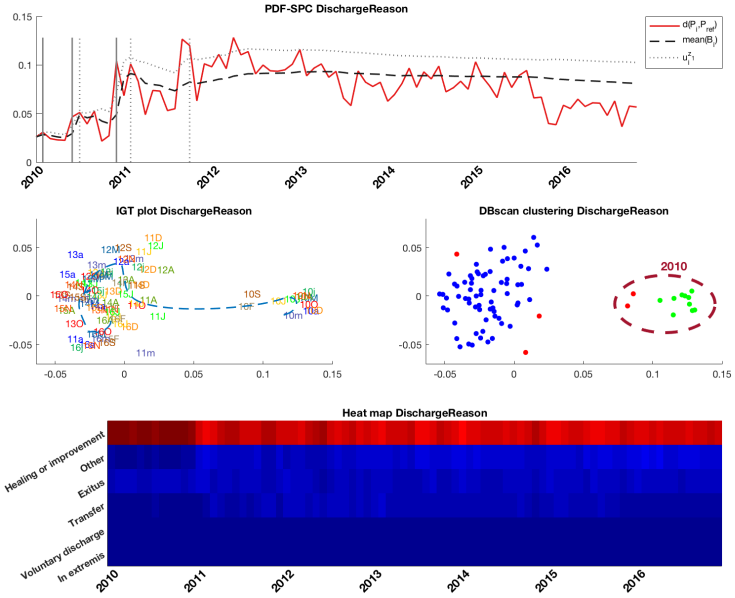


Fig. S1: PDF-SPC, IGT plot, its clustering by DBScan and Heat Map for the variable DischargeReason. An abrupt change is detected at the end of 2010 when the hospital relocation took place. The admittance of patients in delicate health states reduced the number of discharges under “Healing or improvement”.

S2 Figure. Supplementary evidence for I1 and I2.

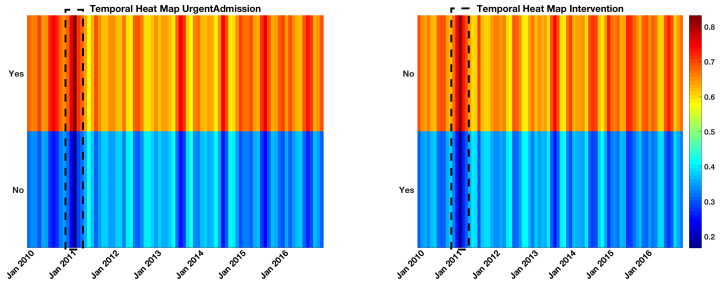


Fig. S2: The color density in February 2011 band shows the increase in the percentage due to the last month relocation. The lower number of observed interventions and an increase in urgent admissions.

S3 Figure. Supplementary evidence for I1 and I2.

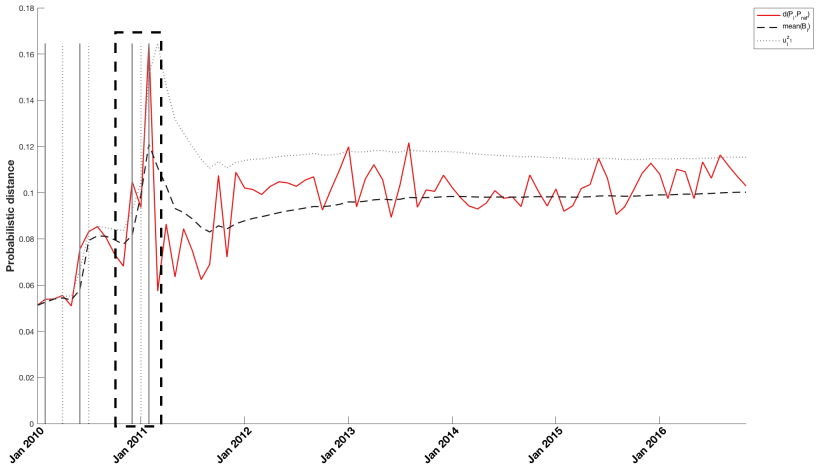


Fig. S3: PDF-SPC of HospitalTransfer. The abrupt changes detected in late 2010 and early 2011 are the results of the hospital relocation.

S4 Figure. Supplementary evidence for I2.

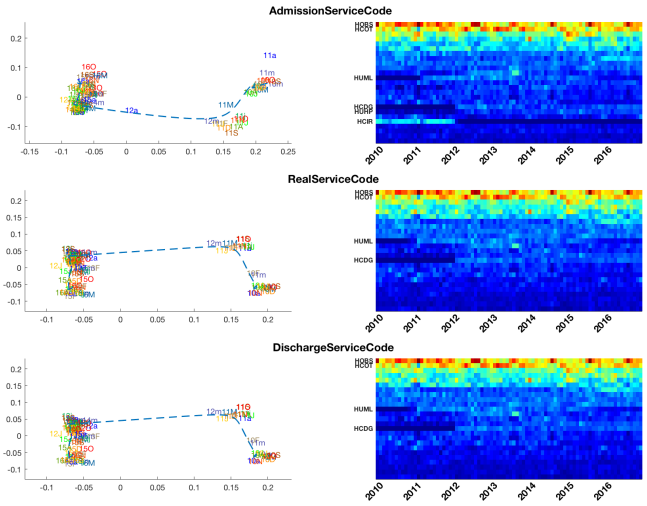


Fig. S4: IGT Plot and Temporal Heat Maps for the Service configurations

S5 Figure. Supplementary evidence for I3.

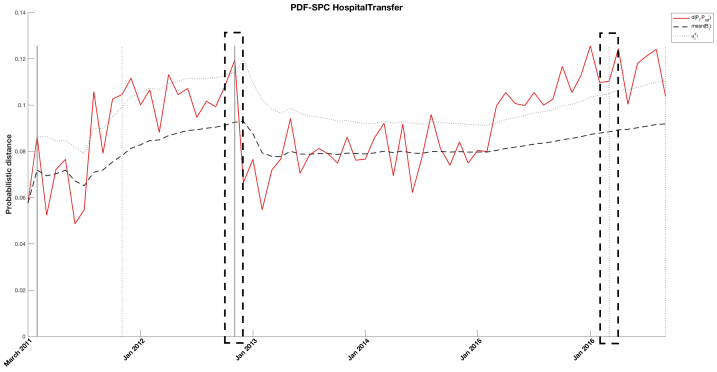


Fig. S5: PDF-SPC for the HospitalTransfer variable. The change caused by a) the opening of the chronic patient's area in the old facilities in early 2013, and b) the new readmittance to the new facilities in early 2016. The changes were detected after removing the cases prior to March 2011 -with the purpose of avoiding the loss of change detection due to the high impact of the hospital relocation-.

S6 Figure. Supplementary evidence for I4.

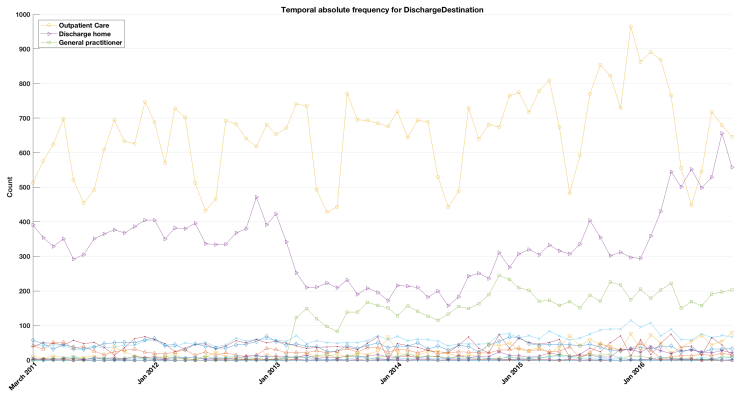


Fig. S6: Temporal absolute count for the discharge destination variable. Since the opening of the chronic service in the old hospital facilities, a new code -in which patients who would be treated in the chronic area were included- was created "Outpatient care". This implies a decrease in the number of patients who were sent home until 2016 when the chronic area was closed.

S7 Figure. Supplementary evidence for I4.

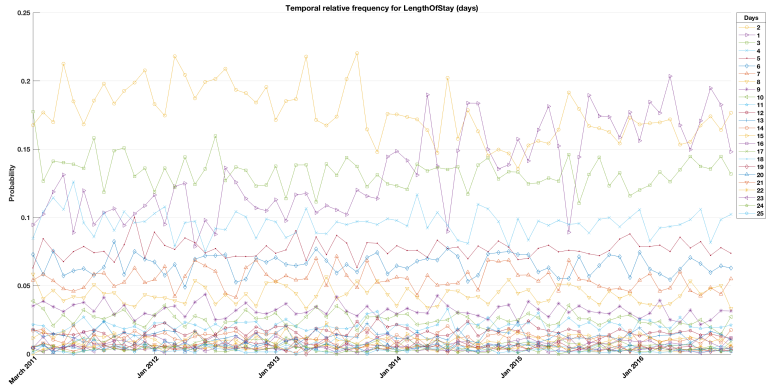


Fig. S7: Temporal relative count for the variable which records the length -in days- of the stay for each hospitalization. The image shows the percentage of 1-day stays increases in detriment of 2-days stays. This shows that the aim of reducing the length of stay, as described in M4, was successful.

4 Journal article (ii)

All our knowledge begins with the senses, proceeds then to the understanding, and ends with reason. There is nothing higher than reason.

Immanuel Kant.

Subgrouping factors influencing migraine intensity in women: A semi-automatic methodology based on Machine Learning and Information Geometry

Pérez-Benito, F.J.[1,2], Conejero, J.A.[2], Sáez, C.[1], García-Gómez, J.M.[1], Navarro-Pardo, E.[3], Florencio, L.L.[4], Fernández-de-las-Peñas, C.[4]

1 Biomedical Data Science Lab. Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
2 Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
3 Departamento de Psicología Evolutiva y de la Educación, Universitat de València, Avenida Blasco Ibáñez, 21, 46010 Valencia, Spain.
4 Department of Physical Therapy, Occupational Therapy, Rehabilitation and Physical Medicine, Universidad Rey Juan Carlos, 28922, Madrid, Spain.

Abstract.

Background: Migraine is a heterogeneous condition with multiple clinical manifestations. Machine-learning algorithms permit the identification of population groups providing analytical advantages over other modeling techniques.

Objective: The aim of this study was to analyze critical features that permit to differentiate subgroups of patients with migraine according to the intensity and frequency of attacks by using machine-learning algorithms.

Methods: Sixty-seven women with migraine participated. Clinical features of migraine, related-disability (MIDAS), anxiety/depressive levels (HADS), anxiety state/trait levels (STAI) and pressure pain thresholds (PPT) over the temporalis, neck, second metacarpal, and tibialis anterior were collected. Physical examination included the flexion-rotation test, cervical range of cervical motion, forward head position in sitting and standing, passive accessory intervertebral movements (PAIVMs) with headache reproduction, and joint positioning sense error. Subgrouping was based on machine-learning algorithms by using Nearest Neighbors algorithms, multisource variability assessment, and Random Forest.

Results: For migraine intensity, group 2 (women with regular migraine headache intensity of 7) were younger, had lower joint positioning sense error in cervical rotation, greater cervical mobility in rotation and flexion, lower flexion-rotation test, positive PAIVMs reproducing migraine, normal PPTs over tibialis anterior, shorter migraine history, and lower cranio-vertebral angle in standing than the remaining migraine intensity subgroups. The most discriminative variable was the flexion-rotation test to the symptomatic side. For migraine frequency, no model was able to identify differences between groups, i.e. patients with episodic or chronic migraine.

Conclusions: A subgroup of women with migraine with common migraine intensity was identify with machine-learning algorithms.

Keywords: Migraine, Random Forest, Machine Learning, Multisource variability

---

4.1 Indroduction

Migraine is a primary headache disorder with a worldwide prevalence of 11.6% within female: male ratio 2:1 [87]. In the last Global Burden of Disease Study, headache (e.g., migraine and tension-type headache) was found to be the second most prevalent pain condition in the world [88]. In fact, health care costs of primary headache in Europe (€13.8 billion) mainly account for migraine and tension-type headache [89].

Migraine attacks are characterized by recurrent episodes of severe headache with accompanying symptoms of autonomic nervous system dysfunction. It is accepted that the pathophysiology of migraine is associated to abnormal neuronal excitability leading to cortical spreading depression and to sensitization of trigemino-vascular pathways [90]. In general, pain is a complex subjective experience that includes sensory-discriminative, affective, and cognitive aspects. In such a scenario, it is usually seen in clinical practice that migraine

can be heterogeneous condition with multiple manifestations. Therefore, the identification of subgroups of patients can help to a better understanding of migraine and provides useful data to support developing clinical decision support systems.

Machine-learning algorithms trained to automatically classify patient populations can be used as classification methods since they provide distinct analytical advantages over other modeling techniques. For instance, supervised machine-learning techniques have the ability to assess all available covariates in every possible clinically meaningful combination and report the combinations in mutually exclusive groups capable of being easily incorporated into decision-support modeling [91]. In fact, they can be combined with network methods for improving prediction and detecting potential correlations between variables [92, 93].

Supervised machine-learning analyses have been able to identify groups of patients experiencing the highest rates of mortality post-interhospital transfer [94]; however, its use is scarce in patients with headache. Garcia-Chimeno et al were able to distinguish with 93% accuracy between patients with sporadic migraine, patients with chronic migraine, and patients at risk of medication overuse via feature selection techniques and machine-learning analyses over diffusion tensor images (DTIs) and questionnaire answers related to emotion and cognition [95]. An overview of how Machine Learning techniques have been used in the general context of pain research has been presented by Lötsch and Ultsch [96].

The intensity and frequency of headache attacks are two features that are clinically used in the differential diagnosis of headaches. For instance, migraine is characterized by headache attacks of moderate-severe intensity lasting 4-72 hours as opposite to headache attacks of mild-moderate intensity lasting from 30 min to 7 days as occurs in tension-type headache [97]. The frequency of headache is mainly used for classification between episodic or chronic headache. The episodic form comprises headache attacks occurring less than 15 days per month, while the chronic comprises headaches occurring 15 or more days/month for more than 3 months and with migraine features on at least 8 days/month [97]. Therefore, we aimed to identify differences in clinical features and the presence of musculoskeletal disorders that permit to subgrouping patients with migraine according to the intensity and frequency of the migraine attacks. We chose these clinical variables for subgrouping since migraine is characterized by moderate-severe intensity of headache and because headache frequency is considered the main outcome in clinical trials. Further, the variables used in this study to subgrouping included clinical features and questionnaires focusing on migraine-related items and also the presence of cervical musculoskeletal impairments, e.g. cervical range of motion, head position, joint position sense error, or reproduction of the headache on manual palpation, commonly associated with primary headaches [98]. We hypothesized that patients with higher intensity and/or higher frequency of

migraine would exhibit more severe musculoskeletal disorders, e.g. lower cervical range of motion, decrease pressure pain thresholds, higher joint position sense error, than those with lower intensity and/or frequency of migraine attacks.

4.2 Methods

4.2.1 Participants

Consecutive women with migraine recruited from a Headache Unit located in a tertiary university-based hospital were included. To be eligible, they had to meet the diagnostic criteria of migraine according to the International Classification of Headache Disorders, 3rd edition [97]. Migraine features including location, years with disease, frequency and intensity of migraine attacks, family history, and medication intake were collected. All participants were screened by an experienced neurologist with more than 20 years of experience in headaches. Participants were excluded if presented any of the following: 1, other primary or secondary headache; 2, history of cervical and/or head trauma; 3, pregnancy; 4, history of cervical herniated disk or cervical osteoarthritis on medical records; 5, underlying systematic medical disease, e.g., rheumatoid arthritis, lupus erythematous; 6, comorbid fibromyalgia syndrome; 7, had received treatment including anesthetic blocks, botulinum toxin or physical therapy within the previous 6 months; or, 8, male gender. All participants signed the informed consent form before their inclusion in the study. The local Ethics Committee of the Hospital Rey Juan Carlos, Spain (HRJ 07/14) approved the study design.

All examinations were held when patients were headache-free and when at least one week had elapsed since the last migraine attack to avoid migraine related allodynia. Since some patients exhibit high frequency of migraine attacks, careful observation of this parameter was considered for examination. If not possible, those women with high frequency of attacks were evaluated at least 48 hours after the last attack. Participants were asked to avoid any analgesic or muscle relaxant 24 hours prior to the examination. No change was made on their prophylactic treatment.

4.2.2 Self-reported Outcomes

A 4-weeks headache diary was used to register clinical features of the migraine [99]: 1, migraine intensity (the mean intensity of the days with migraine attack based on a 11-points Numerical Pain Rate Scale (NPRS); 0: no pain, 10: maximum pain); 2, migraine frequency (days/week); 3, migraine duration (hours/attack).

The Hospital Anxiety and Depression Scale (HADS) was used to evaluate anxiety (HADS-A, 7items) and depressive (HADS-D, 7items) levels [100].

In headache patients, the HADS has shown good internal consistency [101]. Higher scores indicate greater levels of anxiety or depressive levels.

The State-Trait Anxiety Inventory (STAI) was used to assess state (STAI-S) and trait (STAI-T) anxiety levels [102]. The STAI-S assesses relatively enduring symptoms of anxiety at a moment and the STAI-T scale measures a stable propensity to experience anxiety and tendencies to perceive stressful situation as threatening. Both subscales had exhibited good internal consistency and high reliability [103]. Higher scores are indicate of greater state or trait anxiety levels.

The Migraine Disability Assessment Scale (MIDAS) questionnaire was used to assess the degree of related-disability in daily activities (work or school, family and social) caused by migraine [104]. The final score comes from the sum of the missed days regarding the 3 activities.

4.2.3 Widespread Pressure Pain Sensitivity

Pressure pain thresholds (PPTs), i.e., the minimal amount of pressure where a sensation of pressure first changes to pain, were bilaterally assessed with an electronic algometer (Somedic AB, Farsta, Sweden) over the temporalis muscle, the cervical spine, the second metacarpal and the tibialis anterior muscle following previous guidelines [105]. All participants attended a session for familiarization with the pressure test procedure over the wrist extensors. The order of assessment was randomized. The mean of 3 trials on each point was calculated and used for the analysis. Since no side-to-side differences were observed, mean of both sides were used in the analysis. Participants were asked to avoid any analgesic or muscle relaxant 24 hours prior to the examination.

4.2.4 Physical Examination

Physical examination included the musculoskeletal impairments most commonly associated to patients with headache [98, 106]: cervical flexion-rotation test, active range of cervical motion, forward head posture, passive accessory intervertebral movements with head pain reproduction and joint position sense error (JPSE).

The cervical flexion-rotation test (FRT) and active cervical range of motion were assessed as previously described [107]. Briefly, for the FRT, participants were positionednin supine and a CROM® device was placed at their head. The evaluator performed a maximum flexion of the cervical spine followed by rotation toward either side. The rotation limit was determined when the evaluator self-perceived tissue resistance or the patient reported the presence of pain at the upper cervical area. Active cervical range of motion was assessed with a CROM® device and participants seated in a relaxed position on a chair. The CROM® device was positioned on the subject's head and

a familiarization session was performed. The mean of three repetitions was considered in the analysis. This procedure has shown excellent reliability in migraine patients [108].

Forward head position, passive accessory intervertebral movement with headache reproduction and Joint Position Sense Error (JPSE) were assessed following previous guidelines [109]. The cranio-vertebral angle, i.e., the angle between the horizontal plane and a line from the tip of the C7 spinous process to the tragus of the ear, was calculated in sitting and standing positions for assessing forward head posture as previously described [110]. A smaller angle reflects a greater forward head position. Passive accessory inter-vertebral motions were used to evaluate the presence of referred pain to the head elicited by a posterior to anterior (PA) pressure applied to C1-C2 segment in an attempt to provoke a pain response able to reproduce a migraine attack. This procedure has been able to differentiate 3 migraine subtypes: pain-free, local pain, and pain referral to the head [111]. Finally, the JPSE was evaluated by assessing the subject ability to relocate the head to a natural head posture, whilst blindfolded, on active cervical extension, left and right rotations. The difference between the starting (zero) and the position on return was calculated in absolute degrees for each movement tested. Three trials were performed in each direction and the mean JPSE was used in the analysis [109].

All examinations were conducted by an experienced therapist with more than 15 years of experience in the management of headache patients and who was blinded to the migraine headache features (subgrouping classification as described below).

4.2.5 Data Analysis Methods

We considered a fully automated methodology that can be split into 4 steps. Firstly, we first input missing data using the Nearest Neighbors (NN) algorithm. Secondly, we assessed the multisource variability [58, 112]. According to the results, we sub-grouped the variables of migraine intensity and migraine frequency in order to ensure inter-group differences. Finally, random forests classifiers were used to determine physical factors influencing migraine headache intensity and frequency subgroups.

4.2.5.1 Neighbors (NN) algorithm

One of the most widely used algorithms to impute missing data is the NN algorithm. These algorithms are efficient methods to fill in missing data. Each missing value on a record is replaced by a value from related cases in the whole set of records that depends on the type of variable used: categorical missing values are replaced by the mode and quantitative ones are replaced by the mean [113]. The number of neighbors was fixed to 10 before conducting experiments. Several papers including DNA microarray studies [114] (29),

forest inventory [115], or breast cancer [116] have shown benefits of NN as missing data imputer method.

4.2.5.2 Multisource Variability Assessment

This MSV is based on Information Geometry [16, 59], which provide a way for the comparison of dissimilarities between the probability distributions (Probability Density Functions, PDFs) of different data sources. In our case, we modeled headache intensity subgroup distributions using Kernel Density Estimation (KDE) [117]. Due to KDE provides a non-parametric distribution, we used the non-parametric Jensen Shannon distance (JSD) to measure the distance between pairs of PDF's [61, 62]. A JSD is bounded between 0 and 1; where a value of 1 indicates that the compared distributions are disjoint. We constructed a simplex in which each point corresponds to a PDF and each edge joining two points measures the distance between the PDF's. Then, this can be reduced by applying projection methods, such as Principal Component Analysis (PCA) [63] or Multidimensional Scaling (MDS) [65, 118], providing a graphical way to detect inter-group variability.

4.2.5.3 Case labelling

Before conducting the final machine-learning analyses, a preprocess analysis was carried out in the subgrouping variables. The original dataset was completed with two processed variables for grouping, headache intensity and headache frequency due to the low number of cases.

Patients were grouped according to their migraine headache intensity as follows: group 1, patients with migraine pain intensity ranging from 4 to 6; group 2, patients with migraine pain intensity equal to 7 (regular migraine attack pain intensity); group 3, patients with migraine intensity equal to 8; and, groups 4 and 5, patients who suffered headache attacks intensities of 9 and 10, respectively. A second subgrouping according to the frequency of migraine was also identified: group 0, patients with 1 to 8 days per month with migraine (episodic); group 1, patients with 9 to 16 days migraine attacks per month (episodic to chronic); group 2, patients with more than 16 days per month with migraine (chronic).

4.2.5.4 Random Forest Classifier

One of the current trends in machine learning research concerns ensemble methods that combine their results, as the case of Random Forest (RF), which constructs many decision trees that are used to classify by the majority vote [119, 120]. RF classifiers also allow to measure the variables that best explain intra-groups variance. Several authors proved that RF classification outperforms other conventional machine learning algorithms, such as back propagation neural networks and support vector machines and has the

advantages of dealing with unbalanced or multiclass classification problems. These reasons have motivated the use of RF in the current study [121–123].

The parameters were fixed to 512 decision trees composing the forest, the maximum number of decision variables in each tree equal to the $log_2 N$ where N is the number of model inputs and the rest of parameters were fixed to the default proposed by the python implementation of scikit-learn [124].

Due to the number of samples in our database is short, we have used an ensemble of Random Forest to obtain more robust results. Besides, each Random Forest of the ensemble was cross-validated using 8 random stratified folds. This concept consists of creating 8 folds where the proportions of predictor labels are similar to original dataset [125]. A visual description of the ensemble is presented in Figure 4.1. Finally, to assess the performance of the models, the recall and the F1-score were computed [126], according with the equations 4.1. Here, TP_c (True Positive) is the number of patients of a given group c hat are correctly classified, FP_c (False Positive) is the number of patients of other groups that are wrongly classified in the given group c, TN_c (True Negative) is the number of patients of other groups that are not classified in group c, and finally FNc (False Negative) is the number of patients of a given group classified in other groups. The F1-score ranges between $[0, 1]$, being 1 the perfect classification.

$$\begin{aligned} Recall &= \frac{TP_c}{TP_c + FN_c}, \qquad Precision = \frac{TP_c}{TP_c + FP_c} \\ F1 - score &= 2\frac{Recall \cdot Precision}{Precision + Recall} = \frac{2 \cdot TP_c}{2 \cdot TP_c + FP_c + FN_c} \end{aligned} \tag{4.1}$$

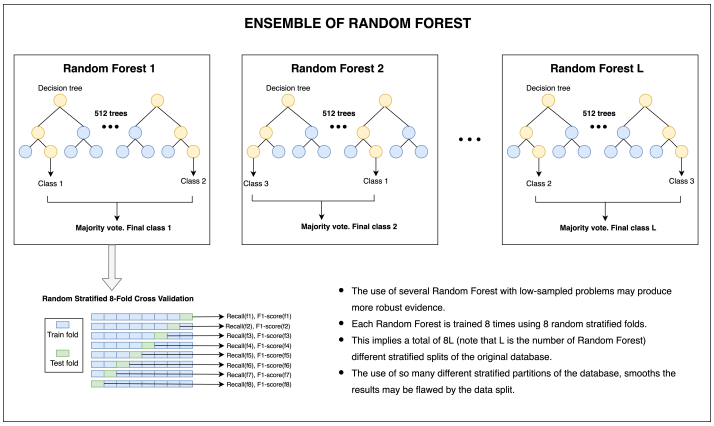


Fig. 4.1: Ensemble of Random Forest. Each Random Forest is composed of 512 decision trees. Each random forest is cross-validated using 8 random stratified folds.

4.3 Results

4.3.1 Participants

Ninety (n=90) consecutive women presenting with headache were screened for eligibility criteria. Twenty-three (25%) were excluded for the following reasons: comorbid headaches (n=10); previous head or neck trauma (n=6); receiving anesthetic block in the past 3 months (n=5) or pregnancy (n=2). Finally, 67 women migraine (20% chronic, mean age: 42 ± 12 years) satisfied all criteria and signed the informed consent. Participants were headache-free at the moment of examination with a mean of 7.5 ± 3.0 days without a migraine attack. Seventy (70%) of the patients self-reported the presence of neck pain mainly during their migraine attacks. Only 4 (6%) self-reported neck pain in interictal phases. Table 4.1 shows clinical, psychological and psychophysical data of the sample.

		Mean (95% CI)
Demographic Features	Age (years)	$42(38-46)$
	History of migraine (years)	$19.8(16.5-23.1)$
Clinical Features	Migraine intensity (NPRS, 0-10)	$8.3\ (7.8-8.8)$
	Migraine duration (hours/attack)	$24.3(19.5-29.1)$
	Migraine frequency (days/month)	$13.0(4.0-21.0)$
	Related-disability (MIDAS)	$45.0(27.5-62.5)$
Psychological variables	HADS-A (0-21)	$12.5(11.5-13.5)$
	HADS-D (0-21)	$10.5(10.0-11.0)$
	STAI-trait (0-60)	$25.7(24.0-27.4)$
	STAI-state (0-60)	$21.7(20.6-22.8)$
PPT (kPa)	Temporalis muscle	$155.0(132.0-178.0)$
	C5-C6 zygapophyseal joint	$131.5(120.0-143.0)$
	Second metacarpal	$190.0(170.0-210.0)$
	Tibialis anterior muscle	$315.0(287.0-343.0)$
Physical Examination	JPSE Extension (degree)	$4.8\ (4.2-5.4)$
	JPSE Cervical Rotation (degree)	$6.0(5.4-6.6)$
	FHP Sitting (CVA, angle)	$35.5(34.0-37.0)$
	FHP Standing (CVA, angle)	$24.0(22.5-25.5)$
	CROM Flexion (degree)	$51.0(47.0-55.0)$
	CROM Extension (degree)	$60.0(56.0-64.0)$
	CROM Latero-Flexion (degree)	$39.0(37.0-41.0)$
	CROM Rotation (degree)	$63.0(60.0-66.0)$

Table 4.1: **Clinical and demographic features of women with migraine**. NPRS: Numerical Pain Rate Scale; MIDAS: Migraine Disability Assessment Scale; HADS-A: Hospital Anxiety and Depression Scale - Anxiety Subscale; HADS-D: Hospital Anxiety and Depression Scale - Depression Subscale; STAI: State-Train Anxiety Inventory; PPT: Pressure Pain Threshold; JPSE: Joint Positioning Sense Error; FHP: Forward Head Posture; CVA: Cranio-Vertebral Angle; CROM: Cervical Range of Motion.

4.3.2 Accuracy of the subgrouping models

After imputing missing data and checking the interclass difference distributions with MSV for migraine intensity (Fig. 4.2A) and frequency (Fig. 4.2B), the dataset was 200 times randomly stratified 8-fold cross-validated. This overcomes the limitation of the low number of individuals. Each of the 200 stratifications produced 8 different folds which contained similar proportions to the original dataset. As can be seen in Table 4.2, the group, to which more patients belong to, has a total of 21 women. Each fold is composed of 2 individuals of this class, and then the number of possible combinations is 210. We chose 200 RF because each of them will be cross-validated using 8 random stratified folds. This gives us a totally of 1600 different splits, which makes almost impossible not to consider the whole set of combinations.

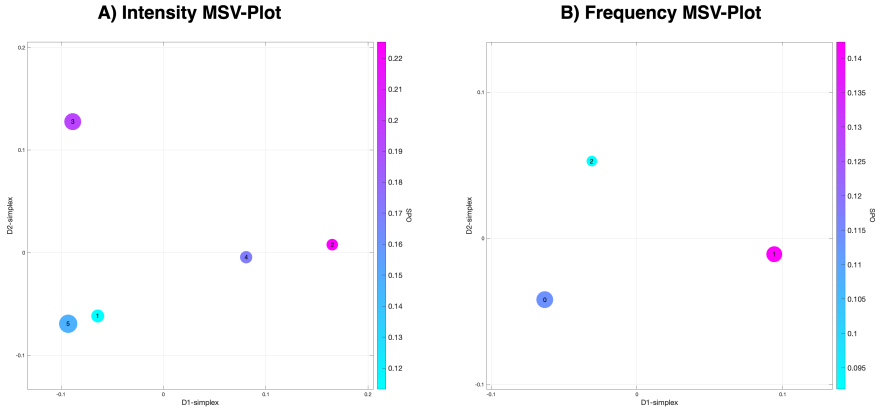


Fig. 4.2: (A) The MSV-Plot for the different intensity classes. (B) The MSV-Plot for the different frequency classes. Source Probabilistis Outlyingness (SPO) measures the Jensen-Shannon distance to the central probabilistic tendency of the whole dataset probability. This metric also ranges between $[0, 1]$. It is worth to mention that distances in B are very small and may not provide enough dissimilarity to be discriminative.

	Group 1	**Group 2**	**Group 3**	**Group 4**	**Group 5**
Frequency total	10	8	17	11	21
Frequency fold	1	1	2	1	2
Frequency total	0.38	0.56	37.28	1.38	67.18

Table 4.2: First row shows the frequency of each group based on migraine intensity subgrouping. Second row shows a typical frequency of each stratified fold, and finally, last row presents the averaged sensitivity for each group.

For migraine intensity, the 8-fold cross-validation averaged recall and frequency of each group are presented in Table 4.2. The averaged F1-score for

the 200 models is shown within Figure 4.3A. Looking at the F1-score, random forest models outperform random classification in a 50% on average. This shows that the variables enclosed in the current study have a certain discriminatory power for determining migraine intensity. The weighted sensitivity mean was 30.86%. It is worth to mention that groups with low density were the worst estimated, because of the low number of cases used to train and to validate the model. Additionally, group 1 contained patients with different headache intensities, which may probably hinder the estimation accuracy. For migraine frequency, the mean accuracy of the 200 implemented models was 0.41, which implied a modest, but not despicable, improvement respect to randomness (Fig. 4.3B). According to the results showed in Table 4.3, none of the random forests was able to find group 2 individuals (a 0 score of sensitivity implies no true positives). This indicates that there was no evidence in the current data which facilitates to discriminate group 2. In this situation, the major possible accuracy score was near to 0.8.

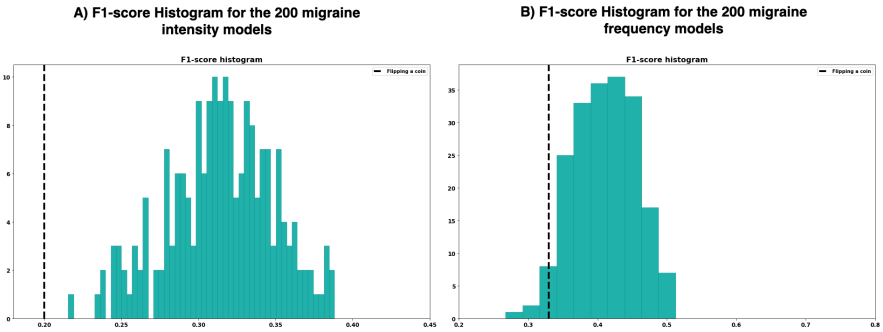


Fig. 4.3: The histogram of the mean F1-score obtained in the 8-fold cross validation of the 200 Random Forest models for migraine intensity (A) and frequency (B) models.

	Group 1	**Group 2**	**Group 3**
Frequency total	30	27	11
Frequency fold	4	4	2
Frequency total	61.51	34.60	0.00

Table 4.3: First row shows the frequency of each group based on migraine frequency subgrouping. Second row shows a typical frequency of each stratified fold, and finally, last row presents the averaged sensitivity for each group. It is worth to mention that the Random Forest based models are not capable to discriminate patients from group 2. It is probably due to the unbalanced samples per class.

An explanation to this fact can be found looking at how random forests models are generated, since they are not robust to unbalanced data and they

usually tend to be biased towards the groups with the majority of elements. Even though the 8-fold cross-validation of the 200 models obtained an F1-score of 0.41 on average, that is a slightly higher than the expected F1-score associated to a random classification, not finding group 2 individuals makes impossible to interpret correctly which variables are influencing the estimation of the migraine frequency.

4.3.3 Variables importance

Random Forests also provide a quantification of the importance of the features within the subgrouping discrimination. The 10 most influential features of each of the 200 models were extracted only for migraine intensity. As it can be seen in Figure 4.4, 20 variables were chosen as the most important from the 200 generated models.

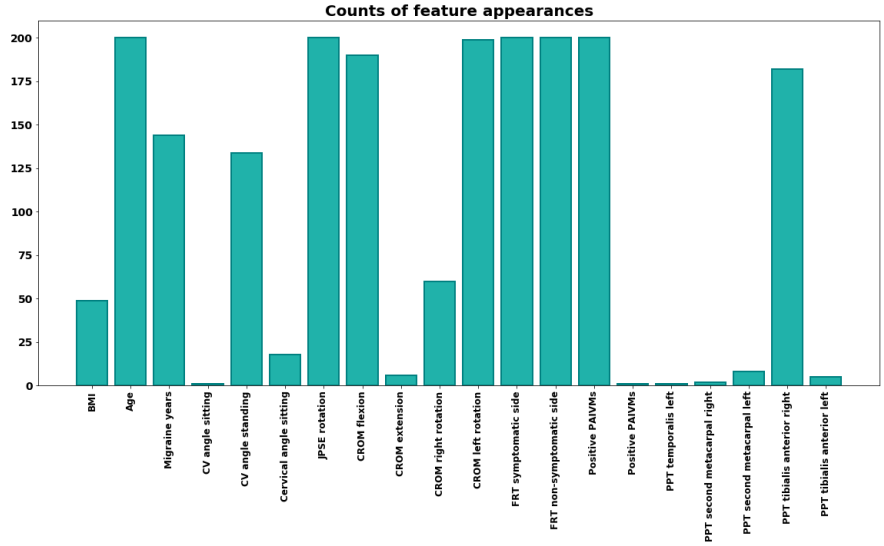


Fig. 4.4: Counting of the variables selected by the RF models. Age, JPSE rotation, FRT symptomatic side, FRT non-symptomatic side and positive PAIVMs were selected as one of the 10 most influential variables by all the models.

For migraine intensity, 6 variables were selected by all the models and other 3 by more than the 50% of the models. Therefore, the results can be considered to be robust. The 10 more frequent variables for identifying subgroup 2 were: age, JPSE in cervical rotation, active cervical range of motion in rotation and flexion, FRT to both symptomatic and non-symptomatic sides, positive PAIVMs, PPT on the tibialis anterior, years with migraine,

and cranio-vertebral angle in standing. In such a scenario, group 2 (women with migraine headache intensity of 7) were younger, had lower JPSE in cervical rotation, greater active cervical range of motion in rotation and flexion, lower FRT to both sides, positive PAIVMs reproducing their migraine headache, normal PPT on tibialis anterior, shorter history with migraine and lower cranio-vertebral angle (i.e., higher forward head posture) in standing position than the remaining groups.

Once these clinical features were selected, we quantify their importance in the discriminative power of the models. In this sense, the histograms of the averaged 8-folds corresponding to each of the 200 models were computed just for migraine intensity (Figure 4.5). The descriptive statistics can be found in Table 4.4. The most discriminative variable in mean over the 200 models after a stratified 8-fold cross-validation was FRT to the symptomatic side (averaged influence of 3.02%).

Variable	**Mean (%)**	**Standard Deviation (%)**
Age (years)	2.59	0.09
JPSE in cervical rotation (degrees)	2.53	0.08
Cervical Range of Motion in rotation (degrees)	2.30	0.07
FRT of the non-symptomatic side (degrees)	2.44	0.08
FRT to the symptomatic side (degrees)*	3.02	0.09
Positive PAIVMs	2.44	0.08
Cervical Range of Motion in flexion (degrees)	2.20	0.07
PPT Tibialis Anterior (kPa)	2.20	0.08
Years with migraine	2.13	0.08
Cranio-Vertebral Angle Standing (degrees)	2.12	0.07

Table 4.4: Descriptive statistics (the percentage of relevance) of the 10 most discriminative variables for migraine intensity. (*) The most discriminative variable for migraine intensity.

4.4 Discussion

A group of women with migraine with common migraine intensity was identified with machine-learning algorithms. Random forest models identified the following most frequent variables in individual trees: age, JPSE in rotation, cervical mobility in rotation and flexion, positive flexion-rotation test, positive PAIVMs reproducing migraine, PPTs over tibialis anterior, migraine history, and cranio-vertebral angle in standing. The most discriminative variable in the model was the flexion-rotation test to the symptomatic side. The random forest model was not able to identify any subgroup depending on the frequency of migraine attacks (episodic, frequent episodic or chronic migraine). These results did not support the a priori hypothesis of this study since individuals with higher intensity or frequency of migraine attacks did not exhibit more severe musculoskeletal disorders.

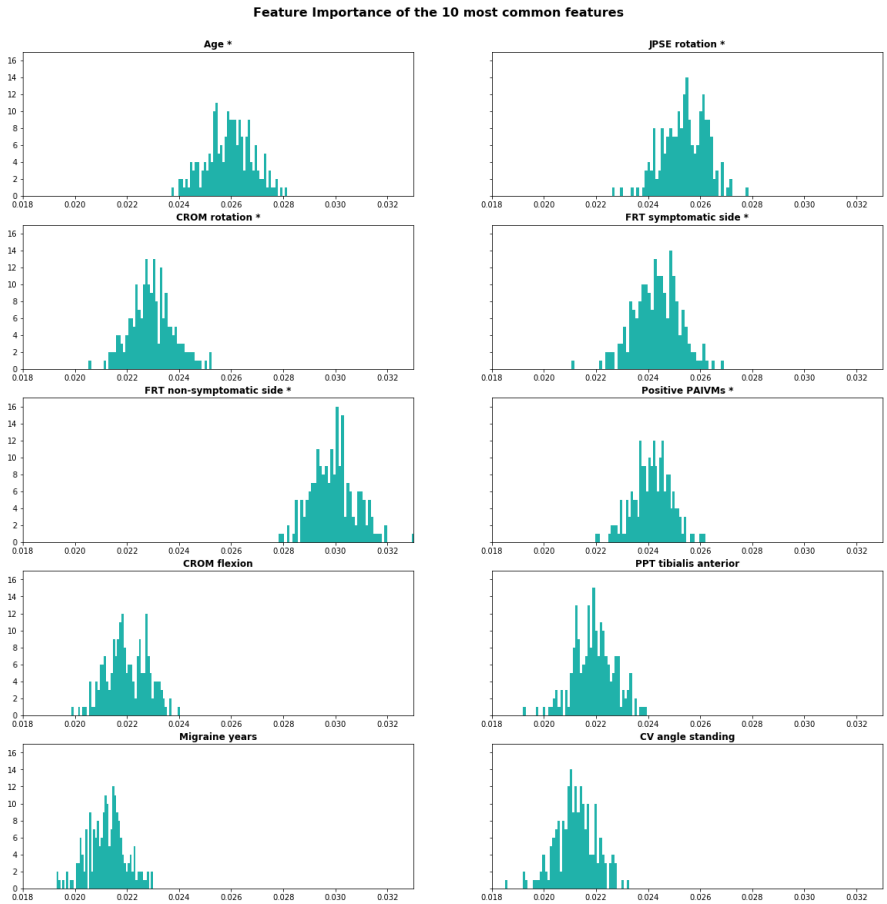


Fig. 4.5: The histograms of the importance of the 10 most important variables of the 200 RF models for migraine intensity.

It is important to note that features were selected in the current study to carry out a clinical classification when differentiating groups of women with migraine according to their intensity or frequency of migraine attacks. From a full set comprising clinical, psychological, and psychophysical outcomes and also physical examination a subgroup of women with migraine suffering from pain intensity of 7 (moderate-intense) during their attacks was identified. It is important to note that migraine pain is characterized by headache attacks of moderate-severe intensity lasting 4-72 hours accordingly to the International Classification of Headache Disorders [97]. Since the results were robust, it seems that the random forest classifier model offered an efficient method for classifying this subgroup of migraine sufferers, as it has solid foundations in

terms of statistical learning, enabling to optimize the decision function in the process.

The subgroup of migraine sufferers identified within the random forest model were younger, lower JPSE in cervical rotation, greater cervical mobility in rotation and flexion, lower flexion-rotation test (positive), positive PAIVMs reproducing migraine symptoms, normal PPTs over the tibialis anterior, shorter migraine history, and lower cranio-vertebral angle in standing as compared to other migraine intensity subgroups. The association of these variables with migraine is not new since some previous studies have investigated the presence of cervical musculoskeletal disorders in this population [106–111]; although its association is still questioned. In fact, a recent meta-analysis has concluded that, among several cervical spine musculoskeletal impairments, individuals with migraine exhibit minimally reduced cervical range of motion with no differences in head posture or JPSE as compared to headache-free people [98]. The current study identified a subgroup of women with migraine with some musculoskeletal disorders of the neck, e.g., positive flexion-rotation test, manual examination (PAIVMs) of the upper cervical able to reproduce their migraine symptoms, and greater forward head posture in standing, when compared to other subgroups of women with migraine. Current results agree with some previous studies suggesting a relevant role of the flexion-rotation test [107, 109], the ability of reproducing migraine symptoms with manual examination of the upper cervical spine joints [111] or a forward head position [110] in migraine. In fact, it is interesting to note that other variables identified by the random forest model, such as cervical range of motion or PPTs over tibialis anterior muscle, should not be considered as impaired, since their values were normal. Similarly, shorter migraine history could be also related to the younger age of this group of patients. Therefore, our study identified that subclassification of individuals with migraine is a highly complex process needing sophisticated analysis such as machine-learning algorithms. Additionally, it is probably that musculoskeletal impairments of the cervical spine have different roles, not only, in promoting or precipitating migraine attacks but also in the intensity of the attacks. From a clinical viewpoint, the variables identified in our study would suggest that the upper cervical spine could be more relevant for this subgroup of patients with migraine than in others. This assumption is supported by the fact that this subgroup of patients exhibited normal cervical range of motion but a positive flexion-rotation test, which supports the presence of upper cervical spine impairment. Therefore, examination of musculoskeletal impairments of the cervical spine should focus on specific groups of migraine patients.

We should also discuss that our sample of women with migraine was explored in a headache-free situation for avoiding migraine-related allodynia and other concomitant symptoms. For instance, this situation also permitted the absence of neck pain during our exploration, a common symptom experienced by patients with migraine during their attacks and associated with

a poor clinical presentation [127]. It is possible that patients experienced concomitant neck pain during migraine attacks could also exhibit different musculoskeletal impairments of the cervical spine representing another subgroup.

We were not able to identify by using random forest models a cluster of variables associated with a group of women with migraine according to the frequency of attacks. We used a clinical subgrouping for headache frequency, mostly based on identification of infrequent episodic, frequent episodic, or chronic migraine. The lack of classification based on the frequency of migraine attacks may be related to the fact that some of the outcomes included in our study, e.g., PPTs, [105], active cervical range of motion [108], JPSE [109] or migraine pain reproduction with passive accessory inter-vertebral motion [111], have not been found to be significantly different between individuals with episodic or chronic migraine, whereas the differences in others, e.g., flexion-rotation test [107] are small. It is also possible the small number of patients within the chronic migraine group, as previously reported in the results section, would lead to an unpowered subgrouping. Future studies should investigate variables associated to frequency of migraine attacks with other outcomes, i.e., migraine-related disability, or kinesiophobia.

Finally, although this is the first study using machine-learning algorithms for the identification of groups of patients with migraine, we should recognize some technical limitations. First, we should highlight that the short number of cases in some subgroups, having fewer than 20 subjects/group. This situation could have led to poor classification accuracy due to the dispersion of the decision space, e.g., in the classification according to migraine frequency. Future studies should include larger dataset of patients to avoid this problem and the main goal should bet the percentage of accuracy of the classifier. Second, future studies could include the use of algorithms for feature selection, such as sequential forward/backward floating selection [128], where the dimension of decision spaces would be reduced and therefore the points sparsity. Further, we only included a sample of women with migraine; therefore, current results should not be extrapolated to men with this condition. In addition, the current subclassification was based on clinical findings observed in a headache-free (interictal phase) status; hence, it is possible that examination during an active phase of a migraine attack could lead to different findings.

4.5 Conclusion

A subgroup of women with migraine with common migraine intensity (moderate to intensity, 7/10) was identify by using machine-learning algorithms. The random forest models identified age, JPSE in rotation, cervical mobility in rotation and flexion, positive flexion-rotation test, positive PAIVMs reproducing migraine, PPTs over tibialis anterior, migraine history, and cranio-

vertebral angle in standing as main variables associated with the group of patients. No cluster of variables was identified accordingly the frequency of migraine.

5 Journal article (iii)

All truths are easy to understand once they are discovered; the points is to discover them.

Louis Pasteur.

A deep learning system to obtain the optimal parameters for a threshold-based breast and dense tissue segmentation

Pérez-Benito, F.J.[1], Signol, F.[1], Perez-Cortes, J.C.[1], Fuster-Baggetto, A.[1], Pollán, M.[2,3], Pérez-Gómez, B.[2,3], Salas-Trejo, D.[4,5], Casals, M.[4,5], Martínez, I.[4,5], LLobet, R.[1]

1 Instituto Tecnológico de la Informática, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
2 National Center for Epidemiology, Carlos III Institute of Health, Monforte de Lemos, 5, 28029 Madrid, Spain.
3 Consortium for Biomedical Research in Epidemiology and Public Health (CIBER en Epidemiología y Salud Pública - CIBERESP), Carlos III Institute of Health, Monforte de Lemos, 5, 28029 Madrid, Spain.
4 Valencian Breast Cancer Screening Program, General Directorate of Public Health, Valencia, Spain.
5 Centro Superior de Investigación en Salud Pública CSISP, FISABIO, Valencia, Spain.

Abstract.

Background and Objective: Breast cancer is the most frequent cancer in women. The Spanish healthcare network established population-based screening programs in all Autonomous Communities, where mammograms of asymptomatic women are taken with early diagnosis purposes. Breast density assessed from digital mammograms is a biomarker known to be related to a higher risk to develop breast cancer.

It is thus crucial to provide a reliable method to measure breast density from mammograms. Furthermore the complete automation of this segmentation process is becoming fundamental as the amount of mammograms increases every day. Important challenges are related with the differences in images from different devices and the lack of an objective gold standard.
This paper presents a fully automated framework based on deep learning to estimate the breast density. The framework covers breast detection, pectoral muscle exclusion, and fibroglandular tissue segmentation.

Methods: A multi-center study, composed of 1785 women whose "for presentation" mammograms were segmented by two experienced radiologists. A total of 4992 of the 6680 mammograms were used as training corpus and the remaining (1688) formed the test corpus. This paper presents a histogram normalization step that smoothed the difference between acquisition, a regression architecture that learned segmentation parameters as intrinsic image features and a loss function based on the DICE score.

Results: The results obtained indicate that the level of concordance (DICE score) reached by the two radiologists (0.77) was also achieved by the automated framework when it was compared to the closest breast segmentation from the radiologists. For the acquired with the highest quality device, the DICE score per acquisition device reached 0.84, while the concordance between radiologists was 0.76.

Conclusions: An automatic breast density estimator based on deep learning exhibits similar performance when compared with two experienced radiologists. It suggests that this system could be used to support radiologists to ease its work.

Keywords: Breast density, Entirely Convolutional Neural Network (ECNN), Deep Learning, Dense tissue segmentation, Mammography.

5.1 Background

Mammographic screening is a highly standardized procedure for breast cancer early detection programs, and the acquired mammograms are interpreted by specialized radiologists who batch read up to 50 mammographies per hour [129]. Full Field Digital Mammography (FFDM) is still one of the preferred methods for breast cancer screening programs. Technology innovations provide better imaging features that promote earlier diagnosis of breast cancer.

Percent Density (PD) which measures the percentage of fibroglandular tissue over the total breast, is known to be a marker of breast cancer development risk [130, 131]. The American College of Radiology Breast Imaging Reporting and Data System (BI-RADS) has also reported a breast classification, based on density, shape, and granularity of the dense tissue [132], suggesting that not only the total amount but also its distribution matters [133, 134]. Besides, one of the principal problems in PD assessment is the inter and intra-observer variability [135–138].

In this sense, an automated tool exhibiting a high agreement with several radiologists could serve as one of the first steps in standardizing the read of breast density. Authors of [139] emphasize a human-like automatic tool could be used as fully independent second reader of screening mammograms, where double reading is standard. A second human reader would only arbitrate discrepancies between the first human reader and the system, halving the workload for any screening program where double reading is standard.

Coupled with this are the tremendous opportunities and challenges for research which are brought by healthcare systems [140], in particular, breast screening programs. To manage and model this huge amount of data, the paradigm of Deep Learning (DL) has emerged. The abstraction ability of DL [141] has demonstrated promising results from speech recognition [142, 143], reconstructing brain circuits [144, 145] or predicting the effects of DNA mutations [146, 147] to medical imaging tasks [148, 149].

One of the most widespread paradigms used in computer vision problems solved via DL take advantage of Convolutional Neural Networks (CNN) [150]. It is based on the extraction of features that are of higher-order as the images go through more layers. CNNs are nowadays the state-of-the-art for many recognition and detection tasks [151–153].

The current work presents a fully automated framework for dense tissue segmentation. It includes breast detection, pectoral muscle exclusion and dense tissue segmentation. Among the contributions of this work, we can highlight (1) a preprocessing algorithm dealing with the variability of mammograms acquired from different devices in the training stage, (2) a new regression architecture Entirely CNN (ECNN), whose output are two parameters used as intrinsic segmentation features, improves classical CNN network (3) a loss function which maximizes the DICE score [154] by continuously rebuilding a probabilistic dense tissue mask, and finally, (4) the ability to manually modify the segmentation using the DMScan software [155, 156].

5.2 Methods

5.2.1 Dataset and participants

A multi-center study covered women from 11 hospitals of the *Comunitat Valenciana* which belong to the Spanish breast cancer screening network.

The prior design of the study was a 1:1 case-control to find factors influencing the development of breast cancer. In this sense, a representation of the whole PD spectrum is assured.

The current study contains a total of 1785 women with ages from 45 to 70. For each patient who developed cancer, if available, the contralateral mammogram was taken from the screening visit previous to diagnostic, otherwise, the contralateral mammogram to the one diagnosed with cancer from the most recent screening visit was selected. Finally, if no previous mammogram existed, then the contralateral mammogram at the diagnostic time was extracted. Since in Spain "raw" mammograms are not routinely stored, all the mammograms are of the type "for presentation".

In 10 of the 11 facilities, the cranio-caudal (CC) and medio lateral-oblique (MLO) views were recruited for each woman, meanwhile, the other facility only collected the CC view. A brief summary of data from the different mammography facilities can be found in Table 5.1.

Id	Unit	Mammography device	Number of women	Number of mammograms (Number of reads)
01	Castellón	FUJIFILM	191	382(764)
02	Fuente de San Luis	FUJIFILM	190	380(760)
04	Alcoi	IMS s.r.l. / Giotto IRE (*)	66	132(264)
05	Xàtiva	FUJIFILM	159	318(636)
07	Requena	HOLOGIC / Giotto IRE (*)	28	56(112)
10	Elda	SIEMENS / Giotto IRE (*)	311	622(1244)
11	Elche	FUJIFILM	278	556(1112)
13	Orihuela	FUJIFILM	117	234(468)
18	Denia	IMS s.r.l. / Giotto IRE(*)	38	76(152)
20	Serrería	(**)	177	354(708)
99	Burjassot	Senographe 2000D	230	230(460)
		Total	1785	3340(6680)

Table 5.1: Screening units, their mammography devices and the number of women and mammograms per device. (*) Implies the use of a new device [Gioto IRE] since 2015. (**) The device is not known.

Mammograms were analyzed by two experienced radiologists using DM-Scan [155, 156]. This software provides assisted semiautomatic tools to segment the breast and the fibroglandular tissue and to exclude undesired regions such as pectoral muscle or armpit.

5.2.2 Breast segmentation framework

The segmentation pipeline is composed of a first step covering breast detection and pectoral muscle exclusion, a second step to normalize the histogram variability between acquisition devices, and then, the dense tissue parametric segmentation is carried out using a deep learning model that was trained using an ad-hoc loss function. Details on each of the aforementioned steps are given below.

5.2.2.1 Background and breast detection

We have used a heuristic, iterative algorithm based on connected components to obtain the gray level threshold that distinguishes breast from background. Even though there exist some issues concerning the use of connected components labeling on binary images [157], homogeneous breast shape makes this kind of algorithms suitable to be used for breast segmentation and exhibits perfect breast detection.

The first step of our approach is to assess the histogram of the image. Based on the premise that the most frequent pixel value has to belong to the background, a range of possible breast thresholds is defined.

Then, this range of thresholds is covered until only two homogeneous components are detected. The first step is to assure that the breast is left-oriented and to binarize the image using the first possible threshold, then apply the connected component labeling method. We chose the Scan plus Array-based Union-Find (SAUF) algorithm [158]. Finally, if only two components are obtained, the threshold is set if not, it is continued covering the range of thresholds.

5.2.2.2 Armpit and pectoral muscle exclusion

Several approaches have been proposed in the literature for armpit and pectoral muscle recognition and exclusion. The authors of [159] proposed a method based on homogeneous contours; the work presented in [160] proposed a combination of image processing, genetic algorithm, morphological selection, and polynomial curve fitting. The approach explained in [161] combines fractional differential enhancement methods with iterative thresholding algorithms meanwhile the authors of [162] propose the use of the outputs of three existing algorithms (region growing, thresholding and k-means clustering) as the input of a machine learning-based computer-aided decision system.

The common key observed in all the aforementioned studies is the knowledge that pectoral muscle appears in a triangle of one of the top corners of the image. Based on this premise, we have defined a robust procedure to exclude pectoral muscles founded on negative gradient changes.

After assuring the image is left-oriented, we applied a Gaussian filter and a 50-pixel moving average to smooth edges and remove spurious isolated brightness pixels. As the muscle is a well contrasted border, it tends to be the last remaining after the smoothing process. We iteratively built a polygon that encloses the exclusion area by selecting the pixel with the lowest gradient every 50 rows until the column of the selected pixel was enough close to the left image border. Finally, the vertex that closed the polygon was the first pixel from the top left corner.

5.2.2.3 Normalizing variability between acquisition devices

The pixel size, grey-scale bit resolution, signal to noise ratio or detective quantum efficiency are important concepts related to image quality [163]. The different mammogram acquisition devices show a huge variability in the quality of mammograms. The first experiments carried out produced different performance results depending on the mammography facility. These results influenced the variability assessment among different devices and how it can negatively impact the training of a machine learning model. We evaluated the differences among the histograms of mammograms over the different mammography facilities by applying the framework proposed by Sáez et al. [4, 52] at image level and checking that well-differentiated mammography facility-clusters appeared as can be seen in Figure 5.1a, where the images from medical centers using different devices were extracted.

Mammogram features like resolution or signal to noise ratio depend on the electronic components of acquisition devices and produce a specific signature visible on the image histogram. In this work, we propose a way to standardize them, which leads to better performance when a model using the images of the whole set of the mammography facilities is trained, avoiding the need of a specific model for each acquisition device.

The preprocessing steps proposed are the following, and the comparison of two histograms from two different acquisition devices can be found in Figure 5.1b):

1. Normalize the pixel values of the image between $[0, 1]$.
2. Shift histogram to set the minimum breast tissue pixel to 0.
3. Normalize again the pixel values between $[0, 1]$.
4. Standardize the breast pixel values to a normal distribution $Z \sim N(0, 1)$.
5. Adjust the pixel values so that the mode is 0.
6. Under the assumption that most typical percent density values are below 30% (above 70th percentile) and values under the 30th percentile only belong to fatty tissue, apply a linear stretching from percentile 30 to -1 and from percentile 70 to 1.
7. Apply once more a normalization to ensure inputs for the Deep Neural Network are between $[0, 1]$.

5.2.2.4 Dense tissue segmentation with Entirely Convolutional Neural Network (ECNN)

Recent works address dense tissue segmentation from different points of view. Authors of [164] used a fractal inspired approach and a multiresolution stack representation to extract 3D histogram features, which were used to apply *k-means* [165] to classify each pixel as fatty, semi-fatty, semi-dense or dense.

Another interesting approach is that proposed in [148], in which an unsupervised step to extract features, based on a sparse autoencoder, is followed by a supervised classifier which tried to classify each pixel as pectoral muscle,

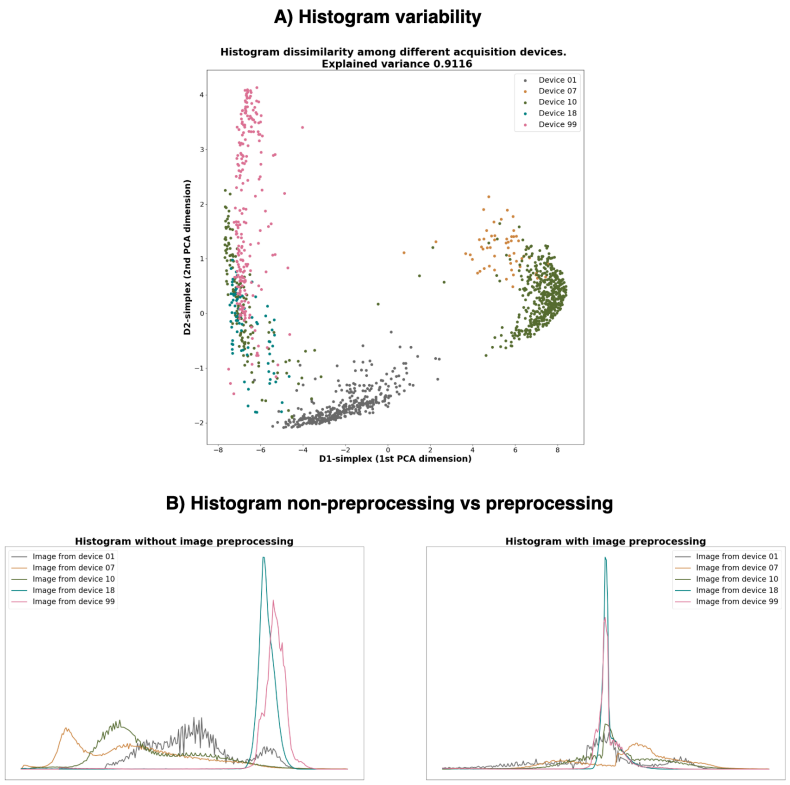


Fig. 5.1: A) Differences among the histograms of the mammograms of the facilities with different acquisition devices. Well-differentiated clusters demonstrated the dissimilarity between acquisition devices. B) Example of histogram transformation using one mammogram from each of the different mammography facilities.

fatty or dense tissue. Close to this approach is the one of [166] that uses 4 fully convolutional networks, two to segment breast tissue on CC and MLO views and the other two to segment the dense tissue on those same views.

Since an accurate and objective *gold standard* does not exist for the segmentation task, the ground-truth of the model to be trained is the segmentation provided by two experienced radiologists who used a semi-atuomatic segmentation tool. Usually, these tools are based on the selection of two thresholds th_B and th_F to segment, respectively, the breast and the fibroglandular tissue. In our study we have used DMScan, a semi-automatic tool that provides a more accurate segmentation using a third parameter α explained below. Therefore, this tool interactively rebuilds a dense tissue mask using the values of three parameters.

- The breast region threshold (th_B). Pixels with values higher than th_B are considered to belong to the breast.
- The brightness corrector α. The X-ray attenuation depends on the thickness of the breast. The thicker the tissue irradiated, the greater the attenuation and, consequently, the brighter the image [155]. The first parameter is related to a brightness correction coefficient k_{ij} by which each pixel is multiplied. The user-defined parameter $\alpha \in [0, 1]$ updates the k_{ij} according to Equation 5.1 where d_{ij} is the horizontal distance of the pixel (i, j) to the image border or the pectoral muscle. It compensates the variation of thickness along the breast.

$$k_{ij} = \alpha + 2(1 - \alpha)d_{ij} \tag{5.1}$$

- The fibroglandular tissue threshold (th_F). Pixels with values higher than th_F are considered to belong to the dense tissue.

We propose an architecture in which convolutions were employed to extract the features needed to replicate the DMScan segmentation as image-intrinsic features: α and th_F. A similar architecture could be applicable to meet the requirements of other semi-automatic threshold-based tools. From now on, we will refer to this architecture as Entirely Convolutional Neural Network (ECNN). It was designed to work with 256×256 px sized images. The proposed architecture and its convolutional layers configuration are shown in Figure 5.2.

Besides, the activation function for the layers was the *Leaky Rectified Linear Unit (ReLU)*, with exception of the last layer which was set to *sigmoid* function to ensure output was $[0, 1]$-bounded. The activation functions are presented in Equation 5.2.

$$ReLU(x) = \begin{cases} x & if x > 0 \\ 0.2x & otherwise \end{cases} \tag{5.2}$$

$$sigmoid(x) = \frac{1}{1+e^{-x}}$$

5.2.2.5 Continuous parameter-based DICE loss function

To measure the performance of our model, we chose the widespread used Sørensen-Dice Similarity Coefficient [154] which measures how much two masks M_1 and M_2 overlap according to equation 5.3.

$$DICE(M_1, M_2) = \frac{2|M_1 \cap M_2|}{|M_1| + |M_2|} \tag{5.3}$$

The use of mean squared error is not monotonically related to the DICE score, leading to an erratic convergence on the learning stage. Furthermore, DICE is the function we want to maximize as it measures the agreement

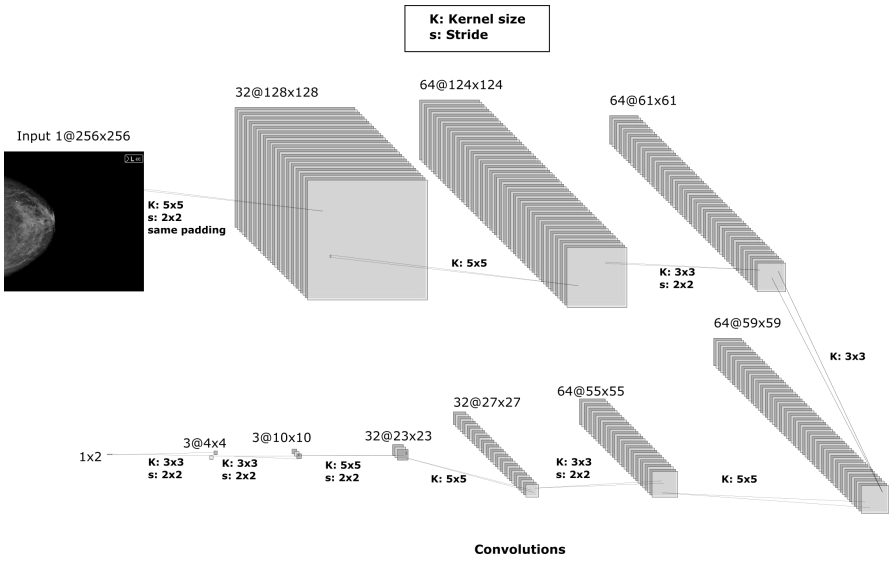


Fig. 5.2: **Entirely Convolutional Neural Network (ECNN) architecture.** The kernel and the strides size for each layer are shown, padding was added to the first convolution to preserve information on the borders. Only convolutions are used to extract the features (α and th_F) needed to segment the dense tissue.

between binary masks. Maximizing DICE is equivalent to minimizing $1-$ DICE. Given two masks M_1 and M_2, a DICE of $\frac{2}{3} = 0.66$ means that the number of pixels belonging to M_1 and M_2 is equal to the number of pixels that only belong to one of them. A DICE score of 0.8 implies that the number of pixels belonging to only one of the masks half the number of pixels that belong to both masks.

This was the reason to develop our metric based on DICE to be used as a loss function in the training stage. The underlying key is to build a map of probabilities in which each element represents the probability of the corresponding pixel belonging to dense tissue and, then, apply the DICE score between estimated mask and the dense tissue mask provided by the radiologists (ground truth). The metric can be represented according to Equation 5.4:

$$\begin{gathered}\mathbb{R}^{[0,1]}_{256\times256}\times\mathbb{R}^{[0,1]}\xrightarrow{fil}\mathbb{R}^{[0,1]}_{256\times256}\times\mathbb{R}^{[0,1]}\xrightarrow{logistic}\mathbb{R}^{[0,1]}_{256\times256}\times\mathbb{R}^{\{0,1\}}_{256\times256}\xrightarrow{loss}\mathbb{R}^{[0,1]}\\ fil\Big((m_{ij}),\hat{\alpha}\Big)\longmapsto\Big(\left[\hat{\alpha}+2(1-\hat{\alpha})d_{ij}\right]m_{ij}\Big)\\ logistic\Big((m_{ij}),\hat{th_F}\Big)\longmapsto\left(\tfrac{1}{e^{-(40[m_{ij}-\hat{th_F}])}}\right)\\ loss\Big((m_{ij}),(n_{ij})\Big)\longmapsto 2\tfrac{\sum m_{ij}n_{ij}}{\sum m_{ij}+\sum n_{ij}}\end{gathered}\tag{5.4}$$

Where $m_{ij}\in\mathbb{R}^{[0,1]}_{256\times256}$ is the mammography resized to 256×256 and $n_{ij}\in\mathbb{R}^{\{0,1\}}_{256\times256}$ is the dense tissue mask provided by an specialist. It is worth to mention that in $fil(.)$, d_{ij} is the one defined in Section 5.2.2.4. The logistic function $logistic(.)$ was applied instead of a *step function* to maintain the continuity, and 40 was used as a slope factor to assure a quick transition between 0 and 1.

Finally, the loss function, which from now on will be referred to as Continuous based Parameters DICE loss score (CPDICE) is defined according to Equation 5.5:

$$CPDICE\Big((m_ij),\hat{\alpha},\hat{th_F},(n_ij)\Big)=1-2\frac{\sum\left(1+e^{-40([\hat{\alpha}+2(1-\hat{\alpha})d_{ij}]m_{ij}-\hat{th_F})}\right)^{-1}n_{ij}}{\sum\left(1+e^{-40([\hat{\alpha}+2(1-\hat{\alpha})d_{ij}]m_{ij}-\hat{th_F})}\right)^{-1}+\sum n_{ij}}\tag{5.5}$$

The corpus, consisting of a total of 3340 mammograms and segmented using DMScan by two radiologists (6680 reads), was randomly stratified taking 75% (4992 segmentations) as training set, from which 10% of the segmentations were extracted with validation purposes (*validation set*), and the remaining 25% (1688 segmentations) as test set. Both mammogram reads of the same image were always included in the same set. The maximum number of epochs was fixed to 500, the optimizer for the training stage was the Adam algorithm [25], and finally, the learning rate was set to 0.001.

5.2.2.6 Dense tissue segmentation example

Three examples of ECNN segmentation of test images using the steps previously described can be found in Figure 5.3. The segmentation is compared to those proposed by the two radiologists. The mammograms were recruited using different acquisition devices. The last example shows the emergence of the abdomen that is still not covered by our pipeline and may negatively influence performance results.

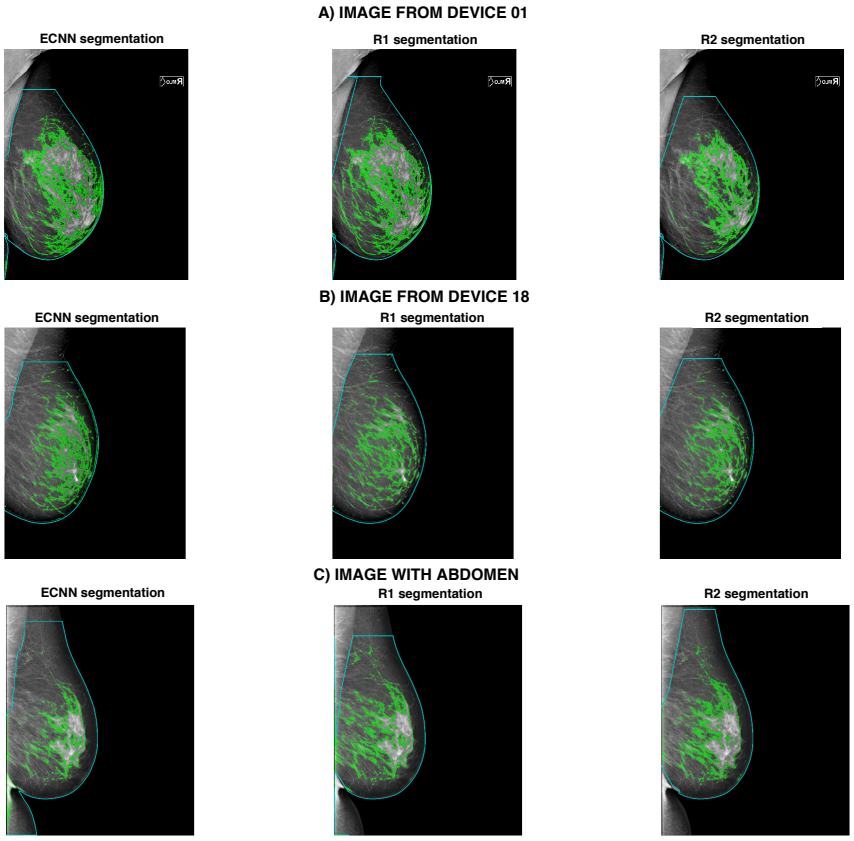


Fig. 5.3: **ECNN segmentation compared to radiologists segmentations on different devices.** A) Segmentation of a mammogram acquired using the device of mammography facility 01. B) Segmentation of a mammogram acquired using the device of mammography facility 18. C) A mammogram from mammography facility 11 where abdomen tissue is found. Medio-lateral oblique mammograms were selected so the exclusion of the pectoral muscle could be seen, however, the abdomen is not excluded.

5.3 Results

As previously mentioned, our model was configured to be trained at most 500 epochs. The lowest loss error obtained was around epoch 400 and the final selected model was then obtained after this number of training iterations.

The lack of a real gold-standard, along with the inter-reader variability [139, 167] motivated us to train our ECNN using segmentations of more than one radiologist as explained before. This decision was made because we did not want a model behaving like a specific specialist, but we wanted

a model that could obtain a level of agreement with any of the specialists comparable to the agreement among them. It is important to note that the segmentation of each radiologist is considered as an independent element. In this sense, if the model gets a perfect segmentation for a mammogram compared to a specific radiologist (R1 for instance), the segmentation of the same mammogram gives a difference concerning the other radiologist (R2) of exactly the difference between R1 and R2. This implies the existence of an unavoidable intrinsic error which has an impact on the performance of the model. It is also worth to mention that radiologists segmentations were labeled using DMScan, which provides an interactive tool to exclude the armpit and pectoral muscle. As can be seen in Figure 5.3, the approach implemented in the current study does not manage, for example, the presence of the abdomen tissue at the bottom of the image. This may also lead to an additional increase of the errors reported in this study.

5.3.1 ECNN as an alternative architecture to standard CNN

As previously mentioned, one of the requirements of the present study is to learn the same parameters that the radiologist has access to. The use of approaches where each pixel or each local region could be freely assigned as dense or not dense was discarded due to the interest in comparing our results with those obtained using widely used threshold-based semi-automatic tools.

Then, to measure the performance of the proposed architecture -ECNN- we trained a fully connected convolutional neural network (CNN) to estimate the desired parameters. A typical architecture for similar tasks [168] composed of a convolutional part followed by a three dense layers (see Table 5.2 for architecture details) provided the intended parameter estimation. It was trained using the CPDICE as a loss function with a learning rate of 0.001.

Layer number	Type layer	Filters/Neurons	Kernel size	Strides	Padding	Activation function
1	Convolutional	32	3×3	1×1	*same*	Leaky ReLu
2	Convolutional	64	3×3	1×1	*valid*	Leaky ReLu
3	Maxpooling	-	2×2	2×2	*valid*	-
4	Convolutional	64	3×3	1×1	*valid*	Leaky ReLu
5	Convolutional	64	3×3	1×1	*valid*	Leaky ReLu
6	Maxpooling	-	2×2	2×2	*valid*	-
7	Dense	512	-	-	-	Leaky ReLu
8	Dense	512	-	-	-	Leaky ReLu
9	Dense	2	-	-	-	Sigmoid

Table 5.2: The details of CNN layers implementation. The first six layers extract image features (convolution stage) and the last three layers play the role of the regressor.

The results per mammography facility compared to those obtained with the ECNN are presented in Table 5.3.

mammography facility	ECNN DICE	ECNN CI	CNN DICE	CNN CI	R1 vs R2 DICE	R1 vs R2 CI
01	**0.81**	[0.78, 0.84]	0.79	[0.76, 0.83]	0.79	[0.76, 0.83]
02	**0.83**	[0.79, 0.86]	0.79	[0.75, 0.83]	0.79	[0.76, 0.82]
04	0.57	[0.50, 0.65]	0.60	[0.53, 0.68]	**0.75**	[0.69, 0.81]
05	**0.84**	[0.81, 0.87]	0.83	[0.80, 0.86]	0.65	[0.61, 0.68]
07	0.85	[0.77, 0.94]	0.81	[0.69, 0.92]	**0.88**	[0.81, 0.96]
10	0.68	[0.65, 0.72]	0.71	[0.67, 0.75]	**0.77**	[0.75, 0.80]
11	**0.87**	[0.85, 0.88]	0.83	[0.81, 0.85]	0.82	[0.80, 0.84]
13	**0.86**	[0.83, 0.89]	0.83	[0.80, 0.87]	0.78	[0.75, 0.82]
18	0.51	[0.40, 0.64]	0.56	[0.46, 0.66]	**0.74**	[0.68, 0.79]
20	0.61	[0.55, 0.67]	0.62	[0.57, 0.67]	**0.78**	[0.75, 0.81]
99	0.78	[0.73, 0.83]	0.75	[0.69, 0.81]	**0.79**	[0.76, 0.82]
Total	**0.77**	[0.75, 0.78]	0.76	[0.74, 0.77]	0.77	[0.75, 0.78]

Table 5.3: ECNN results compared to conventional convolutional architecture. CI refers to 95% confidence interval. ECNN outperforms in many of the devices the agreement between R1 and R2. CNN got better scores on some mammography facilities in which the quality of the mammogram is lower. The DICE scores for the DL models represent the DICE scores to the closer radiologist segmentation.

The conventional convolutional architecture only got significantly better results on mammography facilities 04 and 18. These mammography facilities correspond to the device with the lowest gray-level resolution. The DICE scores in these facilities show also poor agreement between radiologists. Although the best performance of ECNN compared to CNN only can be considered statistically significant for device 11, this approach provided, at least, a similar performance, and it is also faster, more interpretable, and has a lower computational load.

5.3.2 ECNN improvement in function with training epochs

Figure 5.4 shows the model assessment of test images at different epochs (10, 50, 100, 200, 220, 400 and 460) to make clear the achieved balance at different mammography facilities. Averaged-score of validation set also reported its best punctuation at epoch 400 when the validation set score monitored during the training stage.

According to these results, there exist mammography facilities in which the proposed model performance is significantly worse than the obtained in others. It is related to the acquisition device, the quality of acquired images, and probably the unbalanced number of images among different devices.

It should be noted that devices of mammography facilities 1, 2, 5, 11, and 13 come from the same manufacturer and the sum of images in these mammography facilities exceeds by far images coming from other manufacturers. It may influence the good performance at early epochs on images of these mammography facilities. The model seems to improve its results on images from other devices when the local maxima are near to be reached in these

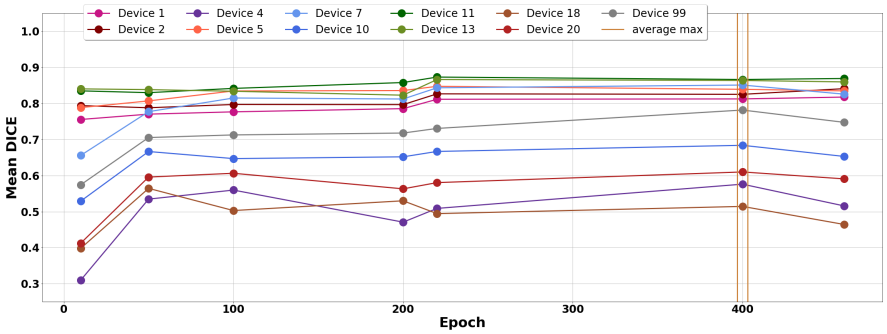


Fig. 5.4: **DICE score per mammography facility at different epochs in the test set.** The first epochs already get acceptable results for images in which the quality is high. As training iterations increase, accuracy increases in these devices and the model is also able to improve its accuracy for the facilities in which their acquisition device image quality is worse. Finally, epoch 400 gets the best averaged score and the model is selected at this point.

mammography facilities which share the same device (the most represented in the corpus).

5.3.3 ECNN segmentation compared with two radiologists

A brief comparison of the obtained DICE scores can be found in Table 5.4.

These results demonstrate a good agreement level of ECNN with segmentations provided by experienced radiologists. As can be seen in Table 5.1, the mammography facilities with a FUJIFILM device (mammography facilities 01, 02, 05, 11, and 13) are those that present better results in Table 5.4. Those mammography facilities presenting lower levels of agreement for the ECNN are also the least populated. This situation makes us suspect that training the model using a balanced number of images per device could increase the reported scores. This probable increment in the performance would be always bounded by the lower gray-level resolution observed in these devices. It also leads to a lower agreement between specialists, with exception of the mammography facility 05 (FUJIFILM acquisition device) where DICE between radiologists is surprisingly low.

Medical facility	test size	ECNN vs closer	R1 vs R2	# ECNN closer to R1 than R2	# ECNN closer to R2 than R1	# ECNN closer to R1 or R2
01	96	0.81	0.79	52	35	58
02	96	0.83	0.79	51	43	63
04	34	0.58	0.75	7	3	8
05	80	0.84	0.65	64	63	76
07	14	0.85	0.88	3	5	6
10	156	0.68	0.77	42	57	65
11	140	0.87	0.82	63	85	100
13	60	0.86	0.78	30	43	49
18	20	0.51	0.74	2	4	6
20	90	0.61	0.78	15	19	27
99	58	0.78	0.79	19	25	35
Total	844	0.77	0.77	348	382	493

Table 5.4: ECNN segmentation DICE scores in function with acquisition devices. Test size column is the number of mammograms available in the test set for each mammography facility. The third column refers to DICE score when ECNN is considered as other radiologist. Fourth column is the DICE score between radiologists. The last three columns show the number of segmentations in which ECNN-R1 are closer than R1-R2, ECNN-R2 are closer than R1-R2 and ECNN-[R1 or R2] is closer than R1-R2.

ECNN outperforms in many devices when compared to the agreement between radiologists and still obtains better results in some devices when it is considered as an specialist. It highlights that almost 60% of ECNN segmentation masks (493 out of 844) are closer to one of the radiologists than the radiologists to each other. This percentage is increased in facilities with FUJIFILM devices. This suggests that ECNN could be considered as an independent reader, but a validation considering the segmentations from other radiologists is needed.

5.3.4 Histogram normalization importance

Figure 5.5 shows how image preprocessing increases the performance of our ECNN.

The substantial increment in the performance of our model, when a preprocessing step is carried out, captures how variability among acquisition devices impacts in the mammogram analysis. These results support the need for standardization of gray-level values from different sources before modeling problems using mammograms.

5.3.5 Specific segmentation model per acquisition device

Having images from different devices could act as a confounder for the models, so the next step was to check if the performance of percent density estimation improved when a specific model is trained for each mammography facility. In this sense, two models using the train images only from one mammography

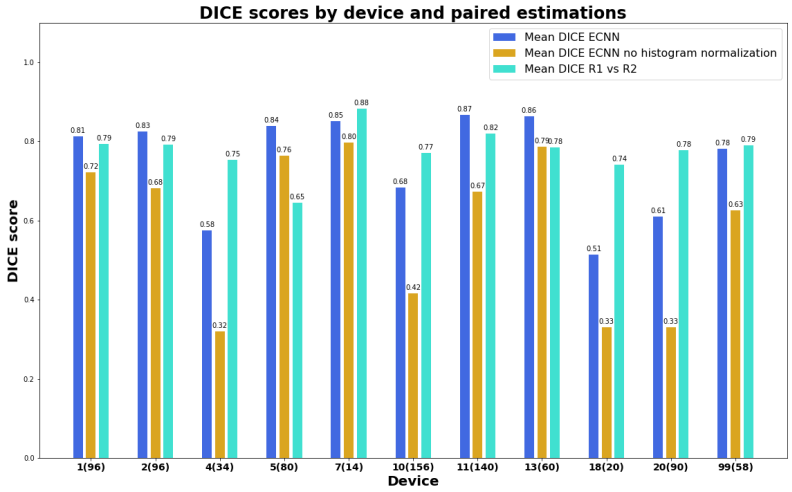


Fig. 5.5: **Comparison of ECNN segmentation using and not using a preprocessing step.** It is observed that results using the proposed histogram normalization outperforms those obtained without any preprocess

facility were trained. One of the models was trained using mammograms from the mammography facility 01 and the other using those from the mammography facility 18. The performance results over the same samples (test corpus from devices 01 and 18) are shown in Table 5.5. They suggest that using a generic model does not imply a substantial loss of performance compared to a specific model.

Medical Center	test size	ECNN vs closer	R1 vs R2	# ECNN closer to R1 than R2	# ECNN closer to R2 than R1	# ECNN closer to R1 or R2
01	96	0.82(0.81)	0.79	41(52)	44(35)	59(58)
18	20	0.58(0.51)	0.74	4(2)	2(4)	5(6)

Table 5.5: Specialized models segmentation DICE scores in function with acquisition devices. Test size column is the number of mammograms available in the test set for each mammography facility. The third column refers to DICE score when ECNN is considered as other radiologist. Fourth column is the DICE score between radiologists. The last three columns show the number of segmentations in which ECNN-R1 are closer than R1-R2, ECNN-R2 are closer than R1-R2 and ECNN-[R1 or R2] is closer than R1-R2. Values in parentheses are the results for the global model.

The specialized model for mammography facility 18 obtained better results when compared to the global model but, still, poor concordance is maintained probably due to the lack of training images and/or the poor quality of them.

5.4 Discussion

According to [139, 169, 170], one of the important tasks for computer-aided diagnosis systems is to provide an accurate and reproducible assessment of mammographic breast density. We consider that our multi-center study demonstrates a good performance of breast density assessment using ECNN, and constitutes a first step in the standardization of how mammographic breast density is assessed. Globally, the score obtained by the proposed framework is comparable, in terms of concordance, to the score obtained by two radiologists.

Typical convolution usage covers pixel-level classification tasks, using convolutional autoencoder architectures [171, 172], or pattern recognition based classification tasks, using fully connected convolutional neural networks [173, 174], or Deep Residual Learning for BI-RADS breast density categories classification [175]. Since our output was continuous, approaches intended to pixel-level classification were discarded. A fully convolutional neural network to estimate the threshold segmentation-based parameters (CNN) was overcome by the architecture in which the desired paramenters are directly extracted as features of the image (ECNN). The performance of the ECNN is better than the obtained by CNN, however this architecture obtain significant better performance for two over the eleven facilities (04 and 18). These mammography facilities have the same acquisition device model and it is also the less represented one in the sample. We expect that increasing the number of images from devices of this model may improve the segmentation results. It is also worth to mention that automatic segmentation applied to the most represented device (FUJIFILM in facilities 01, 02, 05, 11, and 13) were closer to one of the radiologists than each radiologist to the other 73% times (346 out of 472), implying a significant DICE score improvement, outperforming the radiologists concordance.

The main contributions of the present paper can be summarized as:

1. An intuitive preprocess protocol standardizes the histograms of breasts by centering the mode and stretching the tails of the histograms. It allows to extend the range in which the fibroglandular threshold is found. This step reduced the impact of using different acquisition devices.
2. A convolution-based architecture trained to learn the two desired parameters used by radiologists to segment the image. The results provided by this approach obtained slightly better results compared to state-of-the-art algorithms with lower computing workload.
3. An ad hoc, continuous, and differentiable loss function which rebuilds the intended mask from the estimated parameters and assesses the DICE score against the "training ground truth".
4. The approach followed makes easy that a radiologists perform a fine-tuning of the results by interactively modifying the segmentation parameters using a tool such as DMScan.

5.4.1 Limitations and future research

While the parameter based approach was justified to make it compatible with threshold-based semi-automatic tools, exploring other, supervised or unsupervised, mask-based approaches is planned. Supervised mask based approaches could deal with the suboptimal results obtained in some devices and unsupervised approaches would let us complement the models using large databases without the need of human effort.

A second limitation is the pectoral muscle exclusion algorithm. The solution adopted in the present work, although robust, could be improved by taking into account other approaches mentioned in Section 5.2.2.2.

Finally, the use of "for presentation" mammograms instead of "raw" images may be the reason for some of the differences among acquisition devices. It is also desirable to check if "Raw" mammograms would avoid the preprocessing step.

5.5 Conclusion

Nowadays, with the explosion of complex models that can identify features and patterns which are undetectable to the human eye, having a large amount of labeled mammograms is highly necessary for basic and clinical research. In this sense, the availability of a tool that provides automatic segmentation of dense tissue on processed digital mammographies with a high level of concordance with the segmentation of experienced radiologists is desirable.

The work presented in this paper provides an automatic framework based on deep learning which detects the breast, excludes the pectoral muscle, and finally performs a dense tissue segmentation. Our approach is based on the estimation of two segmentation parameters which are learned as image level features. A preprocess step alleviates the influence of the variability among mammograms from different sources and improved the algorithm performance.

The concordance scores (DICE) of the proposed framework are close to the agreement achieved between two radiologists in a multi-center (and multi-device) study. Images from those devices with the highest gray-level resolution provide concordance results even better than those raised by two experienced specialists, suggesting that our model could be used as a fully independent reader. As a final contribution, if the radiologist does not agree with the segmentation proposal, it may easily fine-tuned using a software tool, DMScan, built in our laboratory and freely available for research purposes.

Acknowledgements

The authors of this work like to thank to Guillermo García Colomina, Carlos Barata Ferrando and Empar Giner Ferrando for their support in recruitment and data collection.

Funding

This work was partially funded by Generalitat Valenciana through I+D IVACE (Valencian Institute of Business Competitiviness) and GVA (European Regional Development Fund) supports under the project IMAMCN/2019/1, and by Carlos III Institute of Health under the project DTS15/00080.

Ethics approval and consent to participate

This study was approved by the Research Ethics Committee of the Universitat Politècnica de València (project name: "DM-Scan Herramienta de lectura de densidad mamográfica como fenotipo marcador de riesgo de cáncer de mama") and consent was obtained from study participants at the time of screening.

6 Journal article (iv)

The two enemies of human hapiness
are pain and boredom.

Arthur Schopenhauer.

A happiness degree predictor using the conceptual data structure for deep learning architectures.

Pérez-Benito, F.J.[1], Villacampa-Fernández, P.[2,3], Conejero, J.A.[2], García-Gómez, J.M.[1], Navarro-Pardo, E.[3]

1 Biomedical Data Science Lab. Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
2 Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
3 Departamento de Psicología Evolutiva y de la Educación, Universitat de València, Avenida Blasco Ibáñez, 21, 46010 Valencia, Spain.

Abstract.

Background and Objective: Happiness is a universal fundamental human goal. Since the emergence of Positive Psychology, a major focus in psychological research has been to study the role of certain factors in the prediction of happiness. The conventional methodologies are based on linear relationships, such as the commonly used Multivariate Linear Regression (MLR), which may suffer from the lack of representative capacity to the varied psychological features.Using Deep Neural Networks (DNN), we define a *Happiness Degree Predictor* (H-DP) based on the answers to five psychometric standardized questionnaires.

Methods: A Data-Structure driven architecture for DNNs (D-SDNN) is proposed for defining a HDP in which the network architecture enables the conceptual interpretation of psychological factors associated

to happiness. Four different neural network configurations have been tested, varying the number of neurons and the presence or absence of bias in the hidden layers. Two metrics for evaluating the influence of conceptual dimensions have been defined and computed: one quantifies the influence weight of the conceptual dimension in absolute terms and the other one pinpoints the direction (positive or negative) of the influence.

Materials: A cross-sectional survey targeting non-institutionalized adult population residing in Spain was completed by 823 cases. The total of 111 elements of the survey are grouped by socio-demographic data and by five psychometric scales (Brief COPE Inventory, EPQR-A, GHQ-28, MOS-SSS and SDHS) measuring several psychological factors acting one as the outcome (SDHS) and the four others as predictors.

Results: Our D-SDNN approach provided a better outcome (MSE: $1.46{\cdot}10^{-2}$) than MLR (MSE: $2.30{\cdot}10^{-2}$), hence improving by 37% the predictive accuracy, and allowing to simulate the conceptual structure.

Conclusions: We observe a better performance of Deep Neural Networks (DNN) with respect to traditional methodologies. This demonstrates its capability to capture the conceptual structure for predicting happiness degree through psychological variables assessed by standardized questionnaires. It also permits to estimate the influence of each factor on the outcome without assuming a linear relationship.

Keywords: Deep learning, Data-structure driven deep neural network (D-SDNN), Happiness, Happiness-Degree Predictor (H-DP)

6.1 Introduction

The pursuit of happiness is a universal - both cultural and time wise - core driver of human behaviour. Since ancient times pivotal and referent philosophical figures, as for example Aristotle [1] from West or Zhuangzi [2] from East, devoted much of their work to the idea of happiness as an ultimate purpose of human existence. The major proof that this consciousness pursuit of happiness should be considered as a fundamental human goal is the resolution adopted by the United Nations General Assembly on June 28th, 2012 where March, 20th was proclaimed the International Day of Happiness:

[1] *Happiness depends on ourselves.* Aristotle
[2] *Happiness is the absence of the striving for happiness.* Zhuangzi

> *Recognizing the relevance of happiness and well-being as universal goals and aspirations in the lives of human beings around the world and the importance of their recognition in public policy objectives. Recognizing also the need for a more inclusive, equitable and balanced approach to economic growth that promotes sustainable development, poverty eradication, happiness and the well-being of all peoples [176].*

Consistent with this resolution, the United Nations (UN) has created a civilian based movement for a happier world [177, 178], and took the lead to well-being and happiness as a principal aim in the development and launch of the 17 Sustainable Development Goals of the 2030 Agenda for Sustainable Development [179, 180].

6.1.1 Happiness-Degree Predictor

Since the emergence of Positive Psychology [181] as the scientific study of factors that lead humans – both at the individual and collective level– to thrive, the research community has consistently built up the evidence-based knowledge about the so-called happiness or subjective well-being [182–189].

Happiness and depression are terms employed in daily life to denote affective states and mood swings, which are reliably represented as falling at opposite ends of a bipolar valence continuum [190, 191]. For illustrative purposes, a graphical representation of the emotional valence spectrum is displayed in Figure 6.1.

Fig. 6.1: Emotional valence spectrum

As it can be seen, depression is allocated at the very end of the negative affect side whereas happiness is placed at the opposite one. This implies that happiness is not just the absence of negative mood and affective states, but also the presence of positive ones.

Regarding happiness *predictors*, existent research has found *psychological factors* such as stress coping strategies [192, 193], perceived social support [194–197] or personality [198–201] to have a considerable weight in its emergence. Up to now, the traditional methodological approach employed for happiness degree prediction has been a Multivariate Linear Regression (MLR) [202].

Emerging paradigms, novel approaches, and tools such as deep learning are becoming increasingly influential in psychological research as in the case of

emotion recognition [203–205], sentiment analysis and/or classification [206–208]. It is worth to mention that both topics were endorsed in recent special issues in the last years [209–211] demonstrating the significance of the study and enabling us to avoid one of the pressing constraints of MLR that is the assumption of a linear relationship between the predictors (psychological factors) and the outcome (happiness degree).

Recent studies in sentiment analysis enclosed inside the field of psychology show the tendency to monitor the state of the people through social network activity, image/video and sentence classification [207, 212–214]. These researches show the use of convolutional deep learning approaches which present a better behaviour for feature extraction and selection. Our study aims to mimic –without assuming any linear relationship– the structure of a set of psychometric scales which are conformed by structured data with prediction and interpretation purposes, becoming unnecessary the use of the convolutional technology because of the nature of data.

6.1.2 Motivation of present study

The main objective of our work is to define a *Happiness Degree Predictor* (H-DP) that permits to obtain information of the most significant factors influencing happiness. In particular, this will permit to test the efficiency of increasingly popular regression deep-learning approach in the prediction of Happiness measured in terms of the psychometric *Short Depression-Happiness Scale* (SDHS).

For this purpose, we propose the construction of an intuitive Data-Structure driven Deep Neural Network (D-SDNN) based on the conceptual structure of the psychological factors -emotional distress, personality, stress coping strategies, and perceived social support- for supervised learning. The current technique of deep learning is believed to have many different advantages [214, 215]. Among them, D-SDNN's are expected to improve the correctness of prediction respect to the ones given by MLR, as well as to monitor the influence –weight– that different conceptual dimensions –psychological factors– have in the emergence of a certain degree of happiness and hence in the H-DP.

The rest of the paper is organized as follows. First, in Section 6.2, we provide a short description of the psychometric scales employed to measure the psychological factors used by our D-SDNN. Next, the sample and the data preprocessing procedure are presented. Section 6.3 is devoted to the conceptual scheme and principal features of D-SDNNs. Four D-SDNNs have been trained. Section 6.4 presents our results using a real data and compared to MLR. Impact, contributions, limitations and future work are presented in Section 6.5. Finally, a short conclusion is drawn in Section 6.6.

6.2 Materials

6.2.1 Sample: Issues to consider

Psychological and mental wellbeing has only recently been measurable with valid and reliable measures, but happiness can be understood as satisfaction with life, depression absence, stable extraversion, etc., so even they do not constitute the same construct may be found strong relationships between them. Literature reveals that a lot of sources may influence in happiness, the strongest effects are due to marital status, the relation with the employment, occupational status, leisure and competencies of health and social skills [216]. So, in this paper we have used a specific instrument to assess happiness and we have included other related and different constructs (as coping strategies, personality, emotional distress and social support) in the model in order to design a whole picture of mental and psychological status of the sample.

6.2.1.1 Description of the sample

The target of the cross-sectional survey was the non-institutionalized adult population residing in Valencia. A total of 823 participants completed the survey, 59.8% of whom were women. The mean age was 46 ($\pm$21.1) ranging from 18 to 92 years old. Regarding the educational level of the sample, a 12.2% had not received formal education, 25.8% primary education, 28.7% secondary education, and the remaining 33.3% had received –or were currently receiving– tertiary education. For what it concerns their marital status, 39% of them were single, 41.4% married, 8.3% separated or divorced, and the remaining 11.3% were widow(er).

6.2.1.2 Grounds for exclusion

The sample was collected by 76 different interviewers implying that some of the participants were interviewed by more than one person. We took this fact into account in order to avoid incorrect results. In this sense, if the multiple responses of each repeated participant were equal, then the participant was included, being excluded in the other case.

6.2.2 Descriptions of psychometric scales

Psychometric scales are standardized questionnaires that measure latent variables (psychological factors) through empirical items (behavioral indicators). The procedure of using a psychometric scale comprises a first step where the scale is validated and a second one where its reliability is estimated. In order to be usable, once a scale has been validated in a certain population, its validity does not need to be checked again. However, the reliability of a scale must be checked every time this scale is used over a different sample. There are several indexes to estimate the internal consistency (i.e. reliability)

of a scale. The index most commonly employed is the Cronbach's α coefficient [217]. Therefore, we will present below the different psychometric scales employed in this work to measure latent variables. Cronbach's α coefficients obtained for each scale are presented in Section 6.2.3.

Happiness was measured with the Short Depression-Happiness Scale (SDHS) [191]. It is a 4-point Likert-scale ranging from 0 ("*never*") to 3 ("*often*") with a total of 6 items, 3 of which describe positive feelings (e.g. "*I felt that life was enjoyable*") while three other describe negative feelings –and are hence reverse scored– (e.g. "*I felt cheerless*"). The total score (which may vary between 0 –Depression– and 18 –Happiness–) was computed to obtain the happiness/depression degree for each participant and was employed as gold-standard for supervised-training for the outcome of the D-SDNN.

Coping Strategies are different mental mechanism regarding to manage demands and conflicts and to regulate emotional response and stress. These strategies include the use of personal resources and coping strategies are involved in situations which individuals frequently feel that do not have enough resources or they are not able to answer properly to these demands. Main coping strategies are conductual, cognitive and emotional and could be focussed towards the problem or towards the emotion –that we have at that moment–. Coping Strategies were assessed using the Brief COPE Inventory [218]. It is a 4-point Likert-scale ranging from 1 ("*I usually don't do this at all*") to 4 ("*I usually do this a lot*") with a total of 28 items regrouped in 14 sub-scales of 2 items each: self-distraction, active coping, denial, substance abuse, use of emotional support, use of instrumental support, behavioural disengagement, venting, positive re-framing, planning, humour, religion, and self-blame.

Personality was assessed with the Eysenck Personality Questionnaire Revised-Abbreviated (EPQR-A) [219]. It consists of 4 scales of 6 dichotomous items ("*yes/no*") each that assess neuroticism, extraversion, psychoticism, and sincerity.

Emotional Distress is a feeling that a person or situation is triggering a psychological suffering and could be expresed in different degrees not only cognitive or verbally but through mental or physical symptoms –deppression, anxiety, insomnia, anorexia or poliphagia, upset, vertigo, fatigue, nausea, pain, etc.–. Emotional distress can be interpreted as the opposite status of well-being, happiness, personal satisfaction, welfare, etc. This psychological factor was measured using the 28-item General Health Questionnaire (GHQ-28) [220]. It is a 5-point Likert-scale ranging from 0 ("*not at all*") to 4 ("*much more than usual*") with a total of 28 items regrouped in 4 sub-scales of 7 items each: somatic symptoms, anxiety/insomnia, social dysfunction, and severe depression.

Social Support was assessed with the Medical Outcomes Study (MOS) Social Support Survey (MOS-SSS) [221]. It consists of a first question asking for the number of close friends and close relatives that the person has, plus a

5-point Likert-scale ranging from 1 ("*non of the time*") to 4 ("*all of the time*") with a total of 19 items regrouped into 4 functional support sub-scales of 8, 4, 4, and 3 items per sub-scale. These are: emotional/informational, tangible, affectionate, and positive social interaction.

6.2.3 Descriptions of Data preprocessing

The reliability is referred to the non-systematic error of the measure. It is a feature of the results and can be influenced by the length of the instrument, the homogeneity of the group measured, etc. [222]. The minimum acceptable value of the reliability coefficient depends on the use made of the instrument [223]. In this sense, we first computed the Cronbach's α coefficients for estimating the internal consistency of the psychometric scales in order to check the reliability work prior to use the data gathered with them. The coefficients obtained are summarized in Table 6.1. It is considered an acceptable internal consistency for Cronbach's α for values from 0.70. As it can be seen in Table 6.1, all scales presented a good reliability except for the case of the EPQR-A (that measured personality). Some authors highlight that reliability indices can be influenced by the scale length [224, 225]. Shorter scales usually show lower coefficients than the longer ones, the personality was measured by the abbreviated version of the scale EPQR (the revised scale consist of 100 items while the abbreviated version comprises 24 items) and this may explain the low internal consistency. In any case, we propose the use of the scale but the results regarding this dimension should be interpreted with caution considering the obtained degree of internal consistency.

Psychometric scale	**Cronbach's α coefficient**
SDHS	0.79
Brief COPE Inventory	0.84
EPQR-A	0.42
GHQ-28	0.87
MOS-SSS	0.95

Table 6.1: Cronbach's α coefficients obtained for each psychometric scale

The variables used in this work can be distinguished between numerical or state ones. We pre-processed them differently according to their nature.

State variables (Marital Status and Level of Education) needed recodification before the analysis under the assumption: if two states are related, i.e. exists the possibility of changing from one state to the other, then the codification only differs in one digit, defining an Ordered Binary-Decision Diagram (OBDD) [226] and permiting to use a *dummy* codification [227].

The range of the numerical variables, such as age (discrete data), gender (binary data) and the results of the standardized psychometric scales (continuous data) -including the predictors and the outcome-, are known.

We therefore normalized data for deep neural network's inputs according to equation (6.1), since networks tend to work better when the data are normalized [228].

$$t = (t_{\max} - t_{\min})\frac{x - x_{\min}}{x_{\max} - x_{\min}} + t_{\min}. \tag{6.1}$$

Here, t represents each input variable for the neural network and x the original value for each variable. Note that $x_{\max} - x_{\min}$ and $t_{\max} - t_{\min}$ represent the range of data collected and neural network's inputs, respectively. The use of data in its original range may provoke a need for comparison of the network's output against the real range, in such case:

$$x = x_{\min} + \frac{(t - t_{\min})(x_{\max} - x_{\min})}{t_{\max} - t_{\min}}. \tag{6.2}$$

Values $t_{\max} = 1$ and $t_{\min} = 0$ have been taken in order to use logistic activation function (see (6.3)) in each neuron of the hidden layers.

$$f(x) = \frac{1}{1 + e^{-x}}. \tag{6.3}$$

6.3 Methods

6.3.1 Conceptual scheme

In line with the above objectives mentioned, we have tried to simulate the data conceptual structure in order to gather extra information about the importance of each dimension (i.e. psychological factors) in the H-DP. This architecture can be understood as an ensemble of simpler networks to approximate a function $f : \mathbb{R}^N \longrightarrow \mathbb{R}$. In the context of regression, ensembling some of the neural networks may be better than ensembling all of them [229].

We propose a hierarchical ensembling data driven method for modeling the task in hand. The preconceived data structure has led the layers' ensembling. The items of the psychometric scales employed for measuring the psychological factors used as predictors have been empirically proved to cluster into sub-dimensions and dimensions, i.e. sub-factors and factors [218–221]. We have mimicked this empirically-based conceptual structure in the design of the architecture for our D-SDNN, as it is shown in Table 6.2 and Figure 6.2. We may observe that the 105 inputs included have been regrouped into six main domains:

1 - **Interviewer ID**, which is included in order to control for the influence of the person who was in charge of the data gathering.
2 - **Age, Gender, Marital Status and Level of Education** are Socio-Demographic features and therefore grouped into the conceptual dimension *Socio-Demographic Data*.

3 - **The 28 items from the Brief COPE Inventory** are firstly grouped into fourteen conceptual sub-dimensions: *Active Coping*, *Positive Remaining*, *Acceptance*, *Use of Instrumental Support*, *Self-distraction*, *Religion*, *Self Blame*, *Planning*, *Humour*, *Use of Emotional Support*, *Behavioral disengagement*, *Denial*, *Substance Use* and *Venting*. These are finally grouped into the conceptual dimension *Coping Strategies* that is the psychological factor measured by the *Brief COPE Inventory*.

4 - **The 24 items from the EPQR-A** are firstly grouped into four conceptual sub-dimensions: *Neuroticism*, *Extraversion*, *Psychoticism*, and *Sincerity*; joining together to the conceptual dimension *Personality*, which is the psychological factor that the EPQR-A measures.

5 - **The 28 items from the GHQ** are in the first place grouped into four conceptual sub-dimensions: *Somatic Symptoms*, *Anxiety/Insomnia*, *Social Dysfunction* and *Severe Depression*, which finally conform the conceptual dimension *Emotional Distress*. This is the psychological factor measured by the GHQ-28.

6 - **The 20 items from the MOS-SSS** are firstly grouped into five conceptual sub-dimensions: *Emotional Support*, *Material Assistance*, *Social Relationships* and *Affective Support*. They are joined together to the conceptual dimension *Social Support*, which is the psychological factor that the *MOS-SSS* measures. It should be mentioned that the first item of this scale is related to *the number of friends and relatives you can count on* and this goes directly to the conceptual dimension. Furthermore, this item has been normalized by formula (6.1) taking $x_{\min} = 0$ and $x_{\max}$ the higher value observed in the sample.

PREDICTORS			
Psychometric Scale	**Input/Example of Item**	**Conceptual Sub-dimensions**	**Conceptual Dimensions**
	Interviewer ID	-	-
-	Age, Sex, Marital Status and level of education	-	Socio-Demographic data
Brief COPE Inventory	*"I've been turning to work or other activities to take my mind off things"*	Self distraction Active coping Denial Substance use Use of emotional support Use of instrumental support Behavioural disengagement Venting Positive remaining Planning Humour Acceptance Religion Self Blame	Coping Strategies
EPQR-A	*"Can you easily get some life into a rather dull party?"*	Neuroticism Extraversion Psychoticism Sincerity	Personality
GHQ-28	*"Have you found everything getting on top of you?"*	Somatic Symptoms Anxiety/Insomnia Social Dysfunction Severe Depression	Emotional distress
MOS-SSS	*"Someone to give you good advice about a crisis"*	Emotional Support Material Assistance Social Relationship Affective Support	Social Support

Table 6.2: Data Conceptual Structure. The first two columns correspond to networks' inputs. Columns *Conceptual Sub-dimensions* and *Conceptual Dimensions* are materialised to layers of the deep neural networks as it is shown in Figure 6.2.

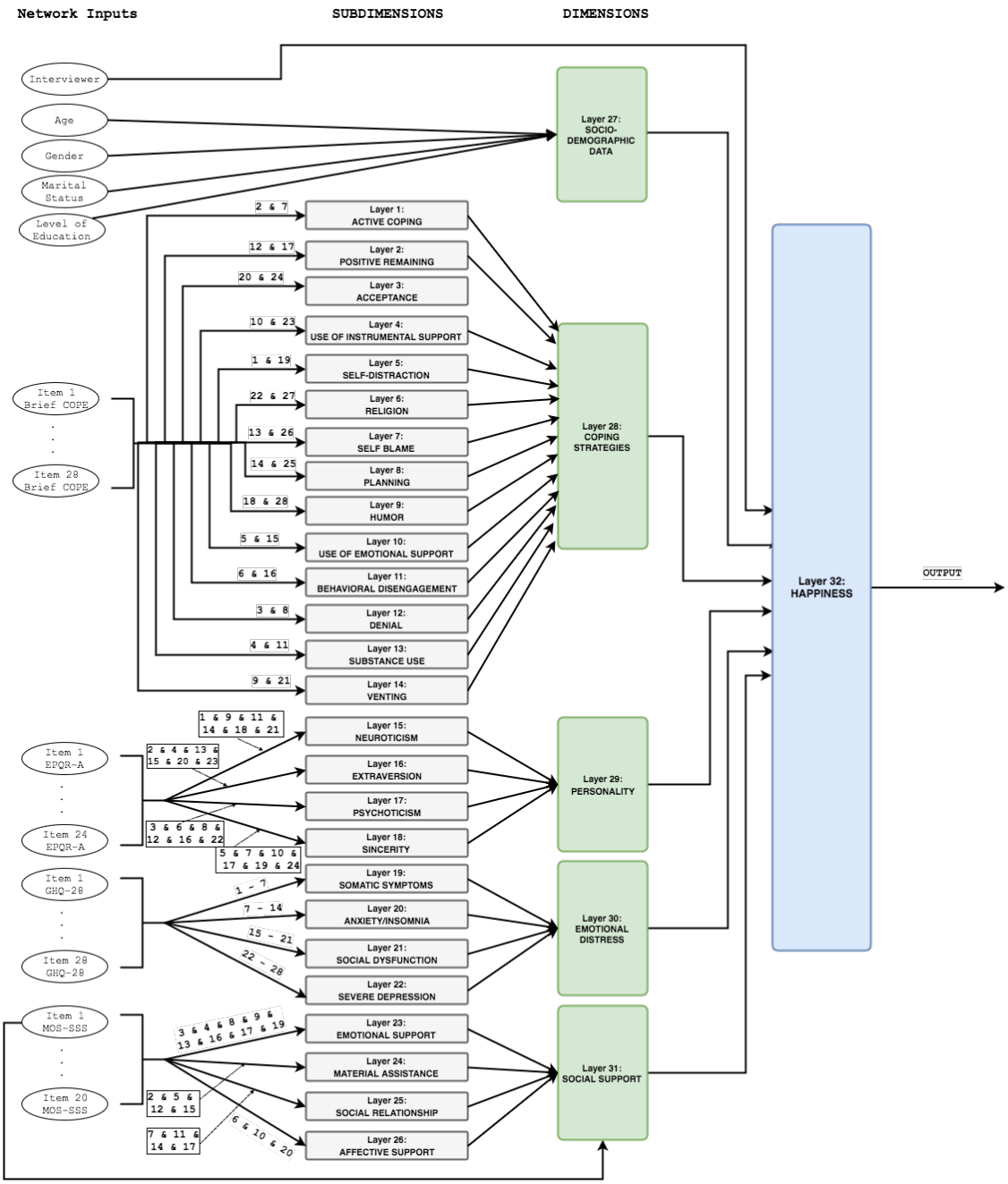


Fig. 6.2: Data-structure driven architecture for our proposed neural networks. The associated number to each arrow, this is arriving to the sub-dimension layers, are related to the number of the items enclosed into the sub-dimension.

6.3.2 D-SDNN features

By mimicking the conceptual structure presented in Figure 6.2, we have created 4 deep neural networks (*net1*, *net1b*, *net2* and *net2b*) for supervised learning, in which each conceptual sub-dimension and dimension conforms one hidden layer.

The four neural networks were the result of combining two conditions with two options in each case:

a) the number of neurons per layer (one vs. as many as incoming inputs), and
b) the *Bias/Variance Dilemma* [230] (existence vs. absence of bias in the hidden layer).

A brief of the configuration of each deep neural network is presented in Table 6.3.

		net1	**net1b**	**net2**	**net2b**
Number of hidden layers		32	32	32	32
Bias in layer		No	Yes	No	Yes
Algorithm for training		L-M	L-M	L-M	L-M
Test for performance		MSE	MSE	MSE	MSE
Initialization algorithm		Random	Random	Nguyen-Widrow	Nguyen-Widrow
	Layers	**net1**	**net1b**	**net2**	**net2b**
Number of Neurons	1 - 14	1	1	2	2
	15 - 18	1	1	6	6
	19 - 22	1	1	7	7
	23	1	1	8	8
	24 - 25	1	1	4	4
	26	1	1	3	3
	27	1	1	4	4
	28	1	1	28	28
	29	1	1	24	24
	30	1	1	28	28
	31	1	1	20	20
	32	1	1	1	1

Table 6.3: Configuration parameters for the tested D-SDNN. Levenberg-Marquardt algorithm for training has been represented as L-M.

In the sequel we will follow this notation: $f(.)$ denotes the logistic function [231] (see (6.3)), x the input vector, w_{ij}^{D} the weight of the i_{th} arriving input into the j_{th} neuron of the conceptual dimension/sub-dimension D, b_h the h_{th} bias vector coordinate and [.] has been used to reflect bias existence or absence depending on the settings of each D-SDNN according to Section 6.3.2. Levenberg-Marquardt has been chosen as training algorithm [232] and MSE as test of performance.

Let $S_1, \ldots, S_{26}$ be the hidden layers that represent the conceptual sub-dimensions of the scales according to Figure 6.2. We denote by $n_{S_1}, \ldots, n_{S_{26}}$

the number of neurons in each layer, $I_{S_1}, \ldots, I_{S_{26}}$ stand for the set of input indexes arriving at each layer with lengths $n^s_{I_{S_1}}, \ldots, n^s_{I_{S_{26}}}$. Then the output of the j_{th} neuron, $j \in 1, \ldots, n_{S_i}$, into the i_{th} sub-dimension layer $\in S_1, \ldots, S_{26}$ is given by

$$s_{ij} = f\left(\sum_{\substack{h=1 \\ l \in I_i}}^{n^s_{I_i}} w^{(i)}_{hj} x_l + [b_h]\right). \tag{6.4}$$

In the same way, let $D_1, \ldots, D_5$ be the hidden layers that represent the conceptual dimension (Socio-Demographic Data, Coping Strategies, Personality, Emotional Distress, and Social Support, respectively). We call $n_{D_1}, \ldots, n_{D_5}$ the number of neurons in each layer $D_1, \ldots, D_5$. Note that the output of the m_{th} neuron in the dimension layer D_1 is given by

$$d_{D_1 m} = f\left(\sum_{i=1}^{4} w^{(D_1)}_{im} x_{i+1} + [b_i]\right). \tag{6.5}$$

For the other dimension layers, the output for the m_{th} neuron in the dimension layer D_k, with $k = 2, \ldots, 5$, being $I_{D_2}, \ldots, I_{D_5}$ the sets of outputs $\{s_{ij}\}$ connected to each layer with lengths $n^d_{I_{D_2}}, \ldots, n^d_{I_{D_5}}$, we have

$$d_{km} = f\left(\sum_{\substack{i=1 \\ t \in I_k}}^{n^d_{I_k}} w^{(k)}_{im} s_t + [b_i]\right). \tag{6.6}$$

We point out that D_5 has an additional connection from one of the inputs (see Figure 6.2) and D_5 must be updated starting from (6.6), x_{86} is corresponding with the first item of MOS-SSS, which is directly connected to the dimension layer as can be shown in Figure 6.2.

$$d_{D_5 m} = d_{D_5 m} + w^{(D_5)}_{(n^d_{I_{D_5}}+1)m} x_{86} + [b_{n_{I_{D_5}}}]. \tag{6.7}$$

Finally, the last hidden layer in all of our proposed schemes of D-SDNN's has only one neuron. The output can be written as

$$y = f\Bigg(w_1 x_1 + \sum_{i=1}^{n_{D_1}} w_{i+1} d_{D_1 i} + \sum_{i=1}^{n_{D_2}} w_{i+1+n_{D_1}} d_{D_2 i} + \sum_{i=1}^{n_{D_3}} w_{i+1+n_{D_1}+n_{D_2}} d_{D_3 i} + \\ + \sum_{i=1}^{n_{D_4}} w_{i+1+n_{D_1}+n_{D_2} n_{D_3}} d_{D_3 i} + \sum_{i=1}^{n_{D_5}} w_{i+1+n_{D_1}+n_{D_2}+n_{D_3}+n_{D_4}} d_{D_5 i} + [b]\Bigg) \in [0, 1]. \tag{6.8}$$

The regression layer (H) provides a value in $[0, 1]$. With (6.2) we can denormalize and obtain values $\hat{y} \in [0, 18]$. Goodness of the fitting will be evaluated according to

$$G_T = \sum_{i=1}^{n_T} \frac{(y_i - \hat{y_i})^2}{n_T}. \tag{6.9}$$

The testing deviation from the original results will be measured according to

$$\delta_t = \sum_{i=1}^{n_t} \frac{(y_i - \hat{y_i})^2}{n_t}, \tag{6.10}$$

where n_T is the training set size and n_t is the testing set size.

Let it be n_{inp} the number of inputs of one neuron of the layer L. In order to measure the global importance of the inputs, we propose the following metrics regarding to weights for each j_{th} neuron in the layer L

$$L_i^{(j)} = \sum_{i=1}^{n_{inp}} \frac{|w_{ij}|}{n_{inp}}, \tag{6.11}$$

and the positivity or negativity of the relationship is determined by

$$\operatorname{sgn}\left(L_i^{(j)}\right) = \operatorname{sgn}\left(\sum_{i=1}^{n_{inp}} w_{ij}\right). \tag{6.12}$$

6.4 Experimental Results

6.4.1 Training, validating and testing the deep neural networks

For each participant we construct a column vector with the inputs for the deep neural network. The first element represents a numeric identifier for the interviewer. From the 2nd to the 5th elements we have the socio-demographic data about the interviewee. The rest of inputs (from the 6th to the 105th) are the responses to the items that conform the standardized psychometric scales.

We have used 578 instances (column vectors) of the total sample, approximately the 70%, for training the 4 tentatives D-SDNNs. Regarding to the other 30%, a 15% has been used for validating and the last 15% for testing.

The fitting with the training data is better for networks with the same number of neurons as incoming inputs (*net2 and net2b*). This implies that we get a better adaptability of multi-neuron layers networks. Besides, within these 2 networks, we can observe that the biased network *learns* so quickly that it falls into over-fitting problems [233]. So as to, these results raise the suspicion that the best network for the database used in the present study is *net2*.

6.4.2 Comparison of D-SDNNs against Multivariate Linear Regression

6.4.2.1 Multivariate Linear Regression

Multivariate Linear Regression (MLR) models are used to predict the value of one or more responses from a set of predictors. MLR's are often used to rate emotional prediction through music [234], effects of colors [235], or neuroimaging [236]. We have constructed a model based on MLR using the inputs of the D-SDNNs as predictors with the purpose of comparison between our D-SDNN's against MLR.

6.4.2.2 D-SDNNs vs MLR

For the construction of the regression model, we proceed in the same way as in Section 6.4.1. We choose the same sample used for training the neural networks proposed (approx. 70%) and we have then calculated the predicted values for the other 245 participants (approx 30%). In the same way, we have evaluated our 4 deep neural networks for the 245 cases excluded of the training set with the purpose of comparison against the same cases predicted by MLR (see Figure 6.3). We have obtained the Mean Square Error (MSE) for each model as shown in Table 6.4.

	MLR	**net1**	**net2**	**net1b**	**net2b**
MSE	$2.30 \cdot 10^{-2}$	$1.54 \cdot 10^{-2}$	$1.46 \cdot 10^{-2}$	$1.58 \cdot 10^{-2}$	$1.86 \cdot 10^{-2}$
Improvement %	0	33	**37**	31	19

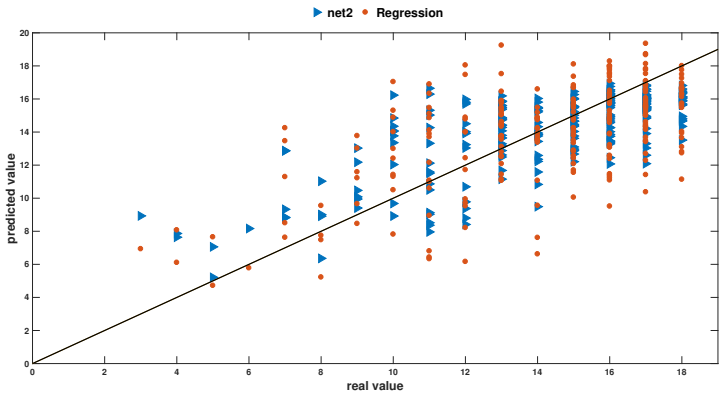
Table 6.4: MSE of the models. The percentage of improvement has been calculated taking as basis MLR. Both observed and predicted values used for the calculus of the MSE were normalized between [0,1] according to (6.1)

.

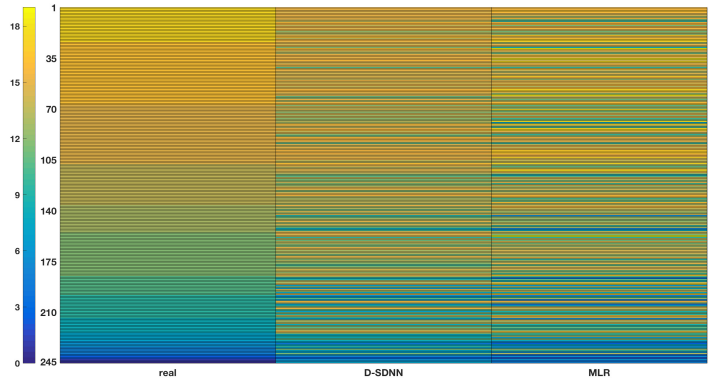
As it can be seen in Table 6.4, the models that best behaved were those generated by deep neural networks. Among them, *net2* stands out, presenting an improvement of 37% taking as basis MLR. It is worth noting here again the significant depletion of MSE in the case of *net2b*. The outstanding performance results of *net2* may be considered as a sign suggesting that the bias added to *net2b* originates an over-training that leads to over-fitting issues causing a detriment in the performance test (i.e. and undermined predictive accuracy).

The predictions obtained by using the D-SDNN *net2* and *Regression* produced a MSE for each possible score as shown in Table 6.5.

It can be observed in Table 6.5 that these scores with more frequency are better predicted by *net2*, i.e. all the scores from 8 to 16 –which represent approximately the 94%– are predicted with more accuracy by *net2*. Besides, those scores that are less frequent present better results for *net2* in cases 5, 6 and 7 improving the percentage of best prediction against MLR up to 97.5%.



(a) net2 vs MLR



(b) Colour Spectrum

Fig. 6.3: Figure (a) presents the comparison of *net2* network against MLR. The points have the value observed as x coordinate, and the predicted value as y. The straight line is $g(x) = x$ which represents the accurate prediction. Figure (b) shows the real, MLR and best fitting D-SDNN color spectrum as indicated in Figure 6.1. Note that MLR color spectrum produces out of range colors.

In the same way, the regression predictions often produce the highest deviations from the expected value, even exceeding the output range (see Figure 6.3). This situation is produced by the little adaptability to data of linear models, which is improved using non-linear methods such as the proposed D-SDNN's in the present study.

SDHS score	Count of cases	MSE net2	MSE Regression
3	1	$4.43 \cdot 10^{-4}$	$1.95 \cdot 10^{-4}$
4	2	$3.55 \cdot 10^{-4}$	$2.65 \cdot 10^{-4}$
5	2	$3.39 \cdot 10^{-5}$	$8.96 \cdot 10^{-5}$
6	2	$5.90 \cdot 10^{-5}$	$6.24 \cdot 10^{-7}$
7	5	10^{-3}	$1.5 \cdot 10^{-3}$
8	4	$1.73 \cdot 10^{-4}$	$1.32 \cdot 10^{-4}$
9	6	$3.87 \cdot 10^{-4}$	$6.46 \cdot 10^{-4}$
10	10	$1.80 \cdot 10^{-3}$	$1.80 \cdot 10^{-3}$
11	16	$1.80 \cdot 10^{-3}$	$2.60 \cdot 10^{-3}$
12	13	$1.20 \cdot 10^{-3}$	$1.90 \cdot 10^{-3}$
13	30	$1.20 \cdot 10^{-3}$	$1.90 \cdot 10^{-3}$
14	19	$7.53 \cdot 10^{-4}$	$1.70 \cdot 10^{-3}$
15	28	$5.37 \cdot 10^{-4}$	$1.40 \cdot 10^{-3}$
16	41	$1.20 \cdot 10^{-3}$	$3.10 \cdot 10^{-3}$
17	43	$1.90 \cdot 10^{-3}$	$3.00 \cdot 10^{-3}$
18	24	$1.80 \cdot 10^{-3}$	$2.80 \cdot 10^{-3}$

Table 6.5: Best model and MLR MSE for each possible score. It is also shown the number of participants who obtained the score. Nobody obtained scores less than 3.

Finally, we have calculated the differences between the values obtained from each prediction model against the ones observed in order to compare the symmetry and the dispersion of the differences. As shown in Figure 6.4, the plot corresponding to the differences between *net2* and the *observed values* is the one with the narrowest box and with the closest outliers.

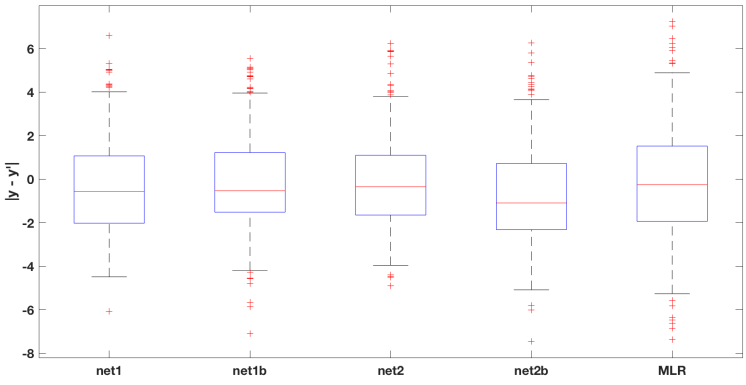


Fig. 6.4: Box-and-whisker plots of differences between predictive models and expected value.

6.4.3 The last layer weights metrics

The weight of each conceptual dimension quantifies its influence in the prediction. Therefore, weights comprehend all the arriving inputs to the last hidden layer. In order to pinpoint the importance that each psychological factor has on the purpose (happiness degree), we have computed metrics (6.11) and (6.12) over each dimension of the best fitting net (*net2*). The results are displayed in Table 6.6. We observed two key values: the weight of the conceptual dimension's influence in absolute terms ($L_{32}^{(l)}$), and the direction of the influence ($sgn\left(L_{32}^{(l)}\right)$).

Accordingly, the most influential dimension in a positive direction for H-DP appeared to be Social Support, whilst the most influential dimension in a negative direction was Coping Strategies. The significantly less influential dimensions were the Interviewer and Socio-demographic Data.

Conceptual dimensions	$L_{32}^{(l)}$	$sgn\left(L_{32}^{(l)}\right)$	**Interpretation**
Interviewer	0.0311	-	Small negative influence
Socio-demographic data	0.1403	+	Small positive influence
Coping Strategies	0.4476	-	Most negatively influential
Personality	0.4186	+	Positively influential
Emotional Distress	0.3897	-	Negatively influential
Social Support	0.5025	+	Most positively influential

Table 6.6: Influence metric values in the best prediction.

6.5 Discussion

6.5.1 Impact

The aim of the present study was the construction of an intuitive D-SDNN based on a set of psychological factors and their sub-components for supervised learning in order to improve traditional methods for H-DP, which are based on linear relationships [237–239]. As expected, when compared with MLR, D-SDNN's show consistent superiority regardless of their configuration (i.e.. number of neurons per layer, and presence or absence of bias). They also allow us to estimate the weight of each psychological factor on the prediction accuracy of the target. According to the best fitting net (*net2*), the psychological factors least influential in the emergence of Happiness were, as expected, the Interviewer and the Socio-demographic data, whereas the most influential ones were Social Support and Coping Strategies. Although the obtained weights might appear weak, they are not. Indeed, for psychological features, it is not only expected to obtain smaller weights than those from artificial devices, but also desirable. This fact prevents people from psychological determinism, i.e. that psychological factors only explain between

a 30 and 50% of the variance allows people to compensate their deficits and to achieve happiness in spite of them.

6.5.2 Contributions

The contributions of the proposed method for H-DP can be summarized in two key points:

(1.1) An intuitive neural network architecture taking advantage of the data conceptual structure which provides the possibility of drawing conclusions about the importance of each conceptual dimension in the outcome measured.

(1.2) Two metrics that allow us to evaluate and quantify the importance of each conceptual dimension on the outcome in absolute terms as well as in which direction (positive or negative).

It is also worth mentioning that the results shown in Section 6.4.1 raised the suspicion that multiple-neuron layer network without bias (*net2*) was the one that would yield better performance because of:

(2.1) It provides enough adaptability to changes within sub-dimensions and dimensions, achieving a better fit to the training dataset.

(2.2) The learning rate was controllable enough to fall into problems of over-fitting what shows the importance of determining under what circumstances it is beneficial or detrimental the use of bias.

After the evaluation and comparison of the chosen test set against MLR (see Section 6.4.2), our results demonstrate a consistently superior performance (for the task in hand) for any neural network. We also point out that:

(3.1) MLR may predict out of range values

(3.2) The best performance for testing set is achieved by *net2*. As shown in Table 6.4.

Using the metrics proposed in (6.11) and (6.12), we have been able to determine the influence of psychological factors in H-DP. The results can be summarized in two main findings:

(4.1) As expected, the people who were in charge of the data collection (i.e. the interviewers) and the socio-demographic characteristics of the participants were the least influential factors for what it concerns H-DP. This means that no matters who asks you, or what your gender, age, marital status or level of education is, your degree of happiness is not likely to be affected.

(4.2) Regarding the role that the studied psychological factors play in the emergence of happiness, we can emphasize:

a. It can be considered congruent with common sense expectations the significantly high and negative influence of Emotional Distress in the degree of happiness.

b. By the same token, it is also consistent with literature the significantly high and positive influence of the perceived Social Support in the degree of happiness. According to these findings, the perceived Social Support may be seen as a buffer for the deleterious effect of the Emotional Distress.
c. The interpretation of the results becomes more controversial for the case of Personality and Coping Strategies. While all the sub-dimensions of the previous factors were in the same direction, is not the case for those of Personality and Coping Strategies (i.e. the influence of some sub-dimensions is expected to be positive, and of some others negative). Concluding that Personality or Coping Strategies, as a whole, have a positive and negative effect, respectively, would very likely be hazardous. One potential explanation is that sub-dimensions, with a positive direction in the case of Personality and with a negative direction in the case of Coping Strategies, have substantially higher weights in absolute terms. However, their respective directions prevail when estimating the general influence of broader dimensions. This would mean that, for example, in the case of Coping Strategies, the adverse effect of Substance abuse or Self blame would be remarkably stronger than the beneficial effect of Humour or Planning.

6.5.3 Limitations

In the case of multi-neuron layer, the proposed metrics for the evaluation of the inputs' influence, eqs. (6.11) and (6.12), can only be conceptually evaluated at the last layer due to the loss of the conceptual scheme within the multi-neuron layers.

By forcing the conceptual structure, the D-SDNNs is not allowed to learn other possible structures that could provide information about the definition of the psychometric scales.

In order to assure the results presented in the present study, the use of the non-abbreviated version of the psychometric scale that measures the personality (EPQR) in the collection of a new data base should be carried out.

6.5.4 Future work

Insights for future works may be arranged in two main points, in order of priority:

(1) In case of multi-neuron layer D-SDNNs, to look for weights characterizations that allow to measure and monitor inputs' influence into each sub-dimension. Besides, it would be interesting to analyze how outputs of the sub-dimensions influence each dimension until reaching the output of the network.

(2) Applying D-SDNN to longitudinal datasets would allow to monitor the variation of weights over time and hence to underpin whether the influence of psychological factors under study changes through the lifespan.

6.6 Conclusions

This paper presented a D-SDNN architecture for H-DP from Socio-Demographic Data and a set of psychological factors (Social Support, Personality, Emotional Distress, and Stress Coping Strategies). The four network configurations used showed better results in comparison with MLR, obtaining an improvement of 37% in the best case.

The best predictor was that employing as many neurons –without bias– as questions endorsed in the sub-dimension or dimension. This prediction obtained a best accuracy in the 97.5% of cases of the population studied in comparison with MLR. It only showed a worst performance –compared to MLR– in SDHS scores with low frequency. The most frequent SDHS score that raised lower MSE for MLR was the value 8 with a relative frequency $\frac{4}{823} \approx 0.4\%$.

Furthermore, this method opens the possibility for conceptual interpretations regarding the importance of each predictor considered: in our study results have shown that socio-demographic characteristics such as gender, age or marital status are not likely to affect the degree of happiness whilst other psychological factors as perceived social support or coping strategies play a major role in the emergence and/or maintenance of happiness.

Based on this, it can be concluded that this study is a new approach of a predictive method, which relies on deep learning architectures by mimicking the conceptual data structure, that presents a consistently superior predictive accuracy together with a better conceptual interpretation.

7 Journal article (v)

To err is human - and to blame it on a computer is even more so.

Robert Orben.

Community detection based deep neural network (CD-DNN) architectures: a fully automated framework based on Likert-scale data

Pérez-Benito, F.J.[1,2], García-Gómez, J.M.[1], Navarro-Pardo, E.[3], Conejero, J.A.[2]

1 Biomedical Data Science Lab. Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas (ITACA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
2 Instituto Universitario de Matemática Pura y Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain.
3 Departamento de Psicología Evolutiva y de la Educación, Universitat de València, Avenida Blasco Ibáñez, 21, 46010 Valencia, Spain.

Abstract.

Deep Neural Networks have emerged as a state-of-the-art tool in very different research fields due to its adaptive power to the decision space since they do not presuppose any linear relationship between data. Some of the main disadvantages of these trending models are that the choice of the network underlying architecture profoundly influences the performance of the model and that the architecture design requires prior knowledge of the field of study.
The use of questionnaires is hugely extended in social/behavioral sciences. The main contribution of this work is to automate the process of a deep neural network architecture design by using an agglomerative hierarchical algorithm that mimics the conceptual structure of such surveys. Although the train had regression purposes, it is easily convertible to deal with classification tasks.

Our proposed methodology will be tested with a database containing socio-demographic data and the responses to five psychometric Likert scales related to the prediction of happiness. These scales have been already used to design a DNN architecture based on the subdimension of the scales. We show that our new network configurations outperform the previous existing DNN architectures.

Keywords: Community detection, Network Science, Deep Learning, Automatic Architecture, Regression, Community-Detection Deep Neural Network (CD-DNN), Happiness, Psychometric scales.

7.1 Introduction

The admissible level of complexity of neural network architectures was constrained until the computational capacity provided by GPU's was unlocked. This new capacity has permitted *Deep Neural Networks* (DNN) to become one of the best tools for classification and regression tasks using data from different science fields [240–242]. Under this new paradigm, DNN's can learn from data represented into subsequent levels of abstraction, which permits to increase its predictive performance [241].

A DNN is composed of several hidden *layers* which, in turn, are divided into neurons. A neuron is composed of a matrix of weights which produces a value after applying an activation function to the entries of the neuron. In general, the design of the network layers structure, the activation function, and the use of bias in the model are human choices, that are usually based on the researcher's knowledge in the problem under consideration. However, sometimes the best architecture that fits a problem can be obtained after a trial-and-error process.

Several attempts of automatic design of DNN architectures have been already considered [243–248] with successful results. Their methods are mainly based on genetic algorithms, hyper-parameter optimization, and reinforcement learning. In all these cases, lots of computational resources are required because of the exponential increment of steps at the training stage.

The study of the network representations of complex processes arisen in physics, biology, and sociology can be represented in terms of interconnected nodes of graphs combining organization and randomness [249, 250]. This field is known as *Network Science*. Some well-known examples include the world-wide-web [251], citation networks [252] the interactome [253] , diseasome [254], or P2P networks [255]. For further information, we refer the reader to Barabasi's book [256]. We will refer to these networks as graphs to avoid confusion with neural networks.

Community detection algorithms permit to decompose graphs into highly inter-connected sub-units [257]. Unlike partitioning, in community detection

procedures the number and size of communities cannot be determined in advance. There is a wide variety of community detection algorithms, covering heuristic [258], divisive [259], and agglomerative algorithms [260, 261].

The current research on graph community detection has many branches. On one side, although almost all community detection algorithms require the global information of the graph, a local approach based on fuzzy relations has been proposed [262]. Furthermore, other approaches based on the trend paradigm of deep learning have emerged, this is known as Graph Neural Networks and has many applications, in particular, community detection. From a Deep Graph Kernels framework that can learn latent representations inside a graph [263] or a variant of a convolutional network that encodes the local graph structure and the features of nodes [264]. Up to adversarial networks that make perturbations on the graph to generate constrained ones which are then classified into communities [265] or the extraction of temporal features using local long short-term memory networks to be used to learn spatio-temporal patterns to infer the communities [266]. Finally, the Graph Attention Networks provide a methodology based on convolutional networks, without depending or knowing the graph structure upfront [267].

In this work, we propose a methodology to design DNN architectures for conceptual-structured data of the answers to items of Likert-type scales. It will rely on the use of community detection algorithms for determining the aggregation of items by similarity in the answers, that is if some items are answered in a similar way (positive or negative) we assume that they will work fine if we consider them in the same layer. To define a DNN with several layers we will apply community detection algorithms at different resolutions. The graph modularity will be a figure of merit of the resulting distribution of the graph into communities. Modularity optimization algorithms belong to the class of agglomerative methods that provide a hierarchical clustering fitting the problem nature. Thereby, methods based on modularity optimization will be our starting point.

The rest of the paper is organized as follows: First, the proposed methodology is presented in Section 7.2. To test the performance of this new method, we apply it to a psychological dataset that has been recently used for predicting happiness [92]. The new proposed DNN architectures enhance the prediction given by the DNN based on the inner structure of psychological scales shown in [92], as it is drawn in Section 7.4 Discussion of the results, including the limitations of the methodology and a proposal for future research lines, are shown in Section 7.5. Finally, some brief conclusions are presented in Section 7.6.

7.2 Proposed Methodology

Our study aims to present a framework to fully automate the development of DNN architectures for solving supervised regression-type problems based

on conceptual-structured data. Our approach that we will call *Community Detection based Deep Neural Networks (CD-DNN)*, consists of:

- Construction of a graph which quantifies the similarity in the answers to items from different scales.
- Application of a community detection method based on modularity optimization at different resolutions.
- Proposal of a DNN architecture automatically inferred from the community hierarchy of the scales items.
- Implementation, training, and validation of the new proposed architecture.

Let us develop each one of these steps.

7.2.1 Construction of a similarity graph from the dataset

Likert scales, or summated rating scales [268], are one of the most commonly used research tools in behavioral sciences. They consist of a list of items that present a finite ordered list of possible answers. The subject who is evaluated answers showing his agreement or disagreement degree with the statement of the item [269]. A value is assigned to each answer, and later all of them are summed. Depending on that value the individual is classified into a group.

Let us suppose that each subject answers to n Likert-type questionnaires, denoted by S_i with $1 \leq i \leq n$, where each questionnaire S_i is composed of n_i items. Let us sequentially rename all the items as $\{v_j : 1 \leq i \leq m\}$, where $m = \sum_{i=1}^{n} m_i$.

Before defining our similarity graph, we recall that a *weighted graph* is given by 3-tuple $G = (V, E, w)$, where V is the set of nodes, $E = \{(v_i, v_j) : v_i, v_j \in V\}$ is the set of edges, and $w : E \to \mathbb{R}_+$ is a function that assign weights to the edges.

Let us considered a dataset in which we have the answers of s_0 subjects to all the scales. Our similarity graph will be a weighted graph $G = (V, E, w)$ defined as follows:

1. We consider $V = \{v_j : 1 \leq i \leq m\}$ as the set of nodes.
2. A pair of nodes is connected by an edge $(v_i, v_j) \in E$ if there exists at least two people answering in the same sense (agreeing or disagreeing) to both items.
3. The weight associated to the edge (v_i, v_j), denoted by w_{ij}, will be the number of subjects answering in the same sense both items v_i and v_j.

7.2.2 Communities detection and architecture construction

Now, we will apply to the similarity graph G community detection algorithms at different resolutions to automatically infer the conceptual hierarchy that will provide us the conceptual hierarchy of the DNN to be trained.

Let us suppose that we have split the graph G into k different communities. The notion of *modularity* of a weighted graph G, denoted by $Q(G)$, was introduced by Newman and Girvan [36] as follows:

$$Q(G) = \frac{1}{2W} \sum_{i,j=1}^{m} \left(w_{ij} - \frac{W_i W_j}{2W} \right) \delta(C_i, C_j) \tag{7.1}$$

where w_{ij} is the weight between nodes i and j, $W_i = \sum_{j=1}^{m} w_{ij}$ is the sum of the weights associated to edges adjacent to v_i, $C_i \in \{1, \ldots, k\}$ is the identifier of the community to which the node v_i belongs to, δ is the Kronecker delta function such that $\delta(u, v)$ is equal to 1 if $u = v$ and 0 otherwise, and $W = \sum_{i,j=1}^{n} w_{ij}$. The value of Q is defined between -1 and 1, and it measures the density of edges inside communities compared to the density of edges between communities.

According to Clauset et al [270], modularity is a property of a graph and a specific proposed division of that graph into communities. The modularity optimization algorithm [36] belongs to the set of agglomerative hierarchical clustering methods [259, 271]. It iterates until a non-null value of modularity is reached. As a reference 0 indicates a random division and 1 the best possible division into communities.

It has been demonstrated that a value above 0.3 is a good indicator of significant community structure in a network [270]. One of the most widespread algorithms for modularity optimization is the Louvain algorithm [37], and it was our choice due to its balance between community detection capability and computation time. Louvain [37] is a 2-phases iterative algorithm which optimizes -in terms of computational time- those proposed by Newman et al. [36] and Clauset et al. [270]. The aim of the algorithm is to find the graph communities which outperform a predefined value of increment of modularity, namely resolution (ΔQ, see Eq. 7.2). This value may vary between 0 and 1. Let us briefly outline how it works: Starting with a weighted graph $G = (V, E, w)$, the first step is to assign a different community to each node $v \in V$. The set of neighbors of a node v_i can be defined as $N(v_i) = \{v_j : (v_i, v_j) \in E\}$.

Let us now consider the gain of modularity (resolution) by moving a vertex v_i into a community C proposed by [37], which in terms of graph can be denoted by:

$$\Delta Q(G, v_i, C) = \left[\frac{W_C + W_{iC}}{2W} - \left(\frac{W_{\hat{C}} + W_i}{2W} \right)^2 \right] - \left[\frac{W_C}{2W} - \left(\frac{W_{\hat{C}}}{2W} \right)^2 - \left(\frac{W_i}{2W} \right)^2 \right] \tag{7.2}$$

where W, W_i are already defined, and

$$W_C = \sum_{v_j, v_k \in C} w_{jk}, \quad W_{iC} = \sum_{v_j \in C} w_{ij}, \quad \text{and } W_{\hat{C}} = \sum_{v_j \in C, v_k \notin C} w_{jk}. \tag{7.3}$$

First, for each node v_i e compute the gains of modularity $\Delta Q(v_i)$, that is assessed by removing v_i from its community and placing it in the community of each one of its neighbours $v_j \in N(v_i)$. Then, we will place v_i in the community for which the gain is positive and maximum. If all the gains are negative, v_i will stay in its original community.

Secondly, we construct a new graph $\widetilde{G}$ whose nodes represent the communities found in the previous step. Two nodes of $\widetilde{G}$ will be connected as long as there was a node in each community that was already connected in G. The weights between two nodes of $\widetilde{G}$ will be given by the sum of the weights of edges between the corresponding two communities. This second step has been shown to preserve modularity of G for $\widetilde{G}$ [272].

Finally, the modularity $Q(\widetilde{G})$ is assessed, and the process is repeated until the resolution reached.

7.3 Proposal of DNN architecture

The next step is to define the DNN architecture. We propose optimizing the modularity at different resolutions ranging between 0.6 and 1 to ensure deviations from randomness.

By optimizing the modularity at different resolutions ranging between 0.6 and 1, we were capable of inferring a hierarchy to develop the DNN architecture automatically. Low-resolution levels produce a higher number of communities while the highest resolution, 1, produces the smallest possible number of them.

To create a DNN architecture we first fix the number of hierarchical levels h_0 in which the layers will be included. We express the range of resolutions, $[0.6, 1]$, in terms of the number of levels according to the next formula, that assigns 0.6 to the first level and 1 to the last one.

$$r_h = 0.6 + \frac{h-1}{h_0 - 1}0.4, \text{ with } 1 \leq h \leq h_0. \tag{7.4}$$

The set of layers that belong to the hierarchical level h will be denoted as H_h. Note that the number of layers in each level is, by definition, the number of communities at resolution r_h.

Let H_1L_j be a layer of the first hierarchical level with $1 \leq j \leq l_1$, and $q(s)$ be one of the inputs of the model with $1 \leq s \leq s_0$, i.e. an answer to the question q by one of the subjects. Then, H_1L_j receives each $q(s)$ as input if, and only if, q belongs to community j at resolution r_1.

Now, let H_iL_j be a layer of the hierarchical level i with $1 \leq j \leq l_i$ and $H_{i+1}L_{j'}$ with $1 \leq j' \leq l_{i+1}$ be a layer of the following level. Then, the outputs of H_iL_j will act as entries for the layer $H_{i+1}L_{j'}$ if, and only if, there exists at least one item in the scale whose community at resolution r_i is j and whose community at resolution r_{i+1} is j'.

Finally, the last hierarchical level only contains one fully connected layer with one neuron, which provides a one-dimensional result which is that optimized using the gold-standard in a supervised learning process.

Now, we have explored the performance of this architecture with different subconfigurations in every layer. We have considered the cases of one neuron per layer, and as many neurons as the number of incoming inputs, that is: for the first hierarchical level it corresponds to the number of incoming items, and for the rest levels it corresponds to the number of incoming layers from the previous one. We have also tested the outcome with and without bias in the hidden layers in both cases.

7.3.1 CD-DNN: Training, validation and test stages

Common practices in machine learning suggest to split the corpus into *Training*, *Test*, and *Validation* sets, to monitor the training process and to avoid over-fitting problems. In this sense, we split the corpus by taking the 70% as the training set, and the remaining 30% as the test set. Half of the test set was extracted as a validation set to monitor the performance between training and validation at each epoch of the training process. So, the training set was composed of the data provided by 578 individuals to the questionnaires. The remaining data was split into the text set (123) and the validation one (122).

We define a heuristic stop rule to avoid over-fitting. This rule is taken as follows:*if during five training epochs the training error was reduced and the validation error was incremented then the training process stops*. Sigmoid function was selected as activation function.

The Adam optimizer [273] was chosen as the training algorithm and the performance test the *Mean Squared Error (MSE)* with the maximum training epoch set to 100. Finally, the results were assessed taking into account the whole test set (the 30% of the original corpus).

7.3.2 Algorithm

The proposed methodology may be summarized in four key points as presented in Section 7.2. Following the notation on this Section, each participant answered m questions.

Firstly, a way to quantify how two questions related was used to define a graph G that represents the similarity between the m questions. Once the graph is built, the number of hierarchical levels needs to be fixed (R). This number lets us choose R thresholds from the resolutions space and apply R times the Blondel's community detection algorithm.

For each resolution, each community defines a layer of the final deep neural network, and two layers of different hierarchical levels are connected if and only if the communities that they represent contain at least one question. The

layers of the first hierarchical level receive as input the questions that belong to the communities it represents. By adding the last layer (fully connected) that receives as inputs the outputs of the layers of the last hierarchical level, the deep neural network is ready to be trained.

The complexity of the graph building is $\mathcal{O}(m^2)$. Due to the definition of the relationship between questions, every pair of nodes have an edge between them. This implies that the complexity of the community detection for each resolution is $\mathcal{O}(m^2)$ but it could be optimized until $\mathcal{O}(m \log(k))$ where k is the average degree [274]. Then, the complexity of the construction of the model is $\mathcal{O}(R \cdot m^2)$ and could become $\mathcal{O}(R \cdot m \log(k))$ where R is the number of hierarchical levels we want to use.

The stages of the procedure may be found in the Algorithm 1. The experimentation for this work was made using *python 3.7* language and the extended libraries *Networkx* [275] and *Tensorflow* [276].

Algorithm 1 The algorithm of the CD-DNN model construction and training.

procedure CD-DNN(d, R) ▷ Building the CD-DNN for the database d with R hierarchical levels
  $d\_train, d\_test = split\_corpus\_train\_and\_test()$
  $d\_validation = random\_half(d\_test)$
  for $0 \leq row < rows(d\_train)$ **do**
    for $row < column < rows(d\_train)$ **do**
      $G \leftarrow obtain\_weight(row, column)$ ▷ Definition of the graph edges.
    end for
  end for
  $m = empty\_model()$
  while $i < R$ **do**
    $r \leftarrow 0.6 + \frac{i-1}{R-1}0.4$
    $C \leftarrow comm\_detection(G, r)$ ▷ $comm\_detection(G, r)$ obtain the Blondel's communities of G at resolution r.
    $inf \leftarrow create\_layers(C)$ ▷ Creates layers for current hierarchical level and its associations with previous level.
    $m.update(inf)$
  end while
  $create\_last\_layer()$ ▷ Creates the last layer. It is dense.
  while $j < 100$ or $not(stop\_criterion)$ **do**
    $model.train(d\_train)$
    $stop\_criterion \leftarrow is\_converging(d\_validation)$
  end while
  return m ▷ m is the trained model
end procedure

7.4 Experimental Results

We will compare the performance of our new automatically designed architectures with the ones in [92], that were applied to the dataset described below.

7.4.1 Dataset

A cross-sectional survey targeting the non-institutionalized adult population residing in Spain was completed by 823 cases. The total of 111 elements of the survey is grouped by socio-demographic data and by five psychometric scales. The psychometric scales measure latent variables describing psychological factors through empirical behavioral indicators. Socio-demographic data covers an identifier of the person who recorded the survey, and the age, gender, marital status and level of education for each individual. The psychometric scales enclosed in the study were:

1. *Short-Depression-Happiness Scale (SDHS)* [191] is a 4-point Likert-scale with a total of 6 items. The total score is a measurement of the happiness degree of the patient (ranging from 0 -Depression- to 18 -Happiness-) and was the gold standard of the model.
2. *Brief COPE Inventory* [218] is a 4-point Likert-scale with a total of 28 items regrouped in 14 sub-scales. The total scale measures coping strategies which are different mental mechanisms regarding manage demands and conflicts and to regulate emotional responses and stress. The 14 subscales represent self-distraction, active coping, denial, substance abuse, use of emotional support, use of instrumental support, behavioral disengagement, venting, positive reframing, planning, humor, religion, and self-blame.
3. *Eysenck Personality Questionnaire Revised-Abbreviated (EPQR-A)* [219] consists of 4 sub-scales of 6 dichotomous items, each that assess neuroticism, extraversion, psychoticism, and sincerity. The total scale measures the personality.
4. *General Health Questionnaire of 28-items (GHQ-28)* [220] is a 5-point Likert-scale assessing the emotional distress which is defined as a feeling that a person or situation is triggering psychological suffering and could be expressed in different degrees not only cognitive or verbally but through mental or physical symptoms. Emotional distress is composed of 4 sub-scales to evaluate somatic symptoms, anxiety/insomnia, social dysfunction, and severe depression.
5. *Medical Outcomes Study Social Support Survey (MOS-SSS)* [221] measures social support. It is composed of 20 items consisting of a first question asking for the number of close people that the person has, plus a total of 19 5-point Likert-scale items that covers 4 functional support sub-scales: emotional/informational, tangible, affectionate, and positive social interaction.

7.4.2 Previous architecture

We will compare the proposed architecture with the one based on the conceptual two-level structure of the scales of the (DS-DNN). We recall that each scale is composed of items. These items are firstly grouped to measure different psychological sub-factors or sub-scales. At the same time, psychological sub-factors are considered altogether to quantify a more general psychological factor. So that, we can consider that the conceptual structure of each scale is composed of two hierarchical levels: the first one to describe psychological sub-scales and the second one to describe the global psychological factor measured by the scale.

We can transfer this structure into a DNN as follows: In the first level of hidden layers, we represent the sub-scales. The items of each sub-scale will be their inputs. For the second hierarchical level the scales as the hidden layers. Here, the outputs of the sub-scales are the inputs of the corresponding scale layers. We also add in this second level a new layer where the socio-demographic data is considered. Its output is directly was included as input in the last layer. This DNN architecture is the same as the one exposed in Section 7.3.

Finally, it is worth to mention that for facilitating the validation of our methodology and to obtain comparable results, the exception of the socio-demographic layer is also considered in the construction of the proposed automatic architecture.

7.4.3 Experimentation

For testing our method with the results obtained with the DNN described in Section 7.4.2, we have automatically generated DNN with 2 hierarchical levels including the socio-demographic layer in the second one. According to Equation 7.4, the resolutions for community detection were 0.6 and 1. The detected communities at these resolutions can be found in Figure 7.1.

The inferred architecture is summarized in Figure 7.2, with a total of 22 layers in the first hierarchical level and 4 layers in the second one, apart from the layer related to socio-demographic information. All the CD-DNN models were trained using the Adam optimizer [273] with random initialization and a learning rate of 0.001 to minimize the MSE. All the layers used the sigmoid function as activation and the last hidden layer, is a dense one which receives as inputs the outputs of all of the layers of the last hierarchical level, in this case, all the outputs of the layers from the second hierarchical level.

Results show that the automatic approach, CD-DNN, outperforms the results obtained by the DNN created using the preconceived survey structure (DS-DNN). The comparison between the results obtained using both methods can be found in Figure 7.3. We highlight that the new proposed methodology tends to be more accurate to predict the less frequent degrees of happiness. This suggests that this methodology has a higher power of abstraction. It is

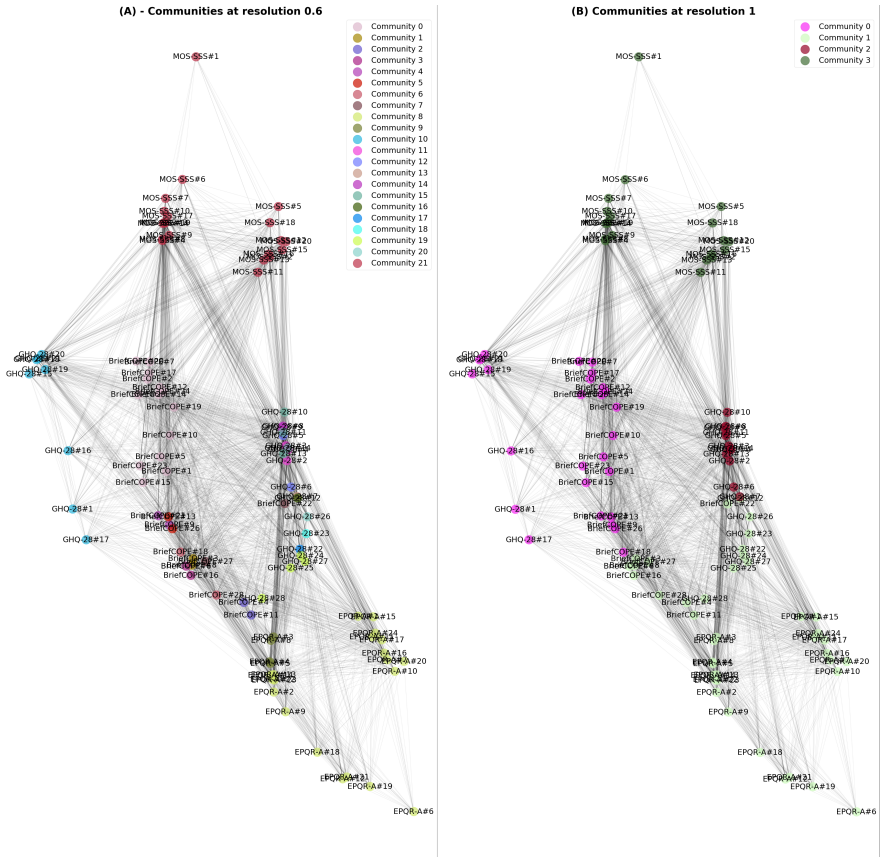


Fig. 7.1: Detected communities after applying Louvain algorithm at different resolutions. Each node represents an item of a psychometric scales. Each node is depicted by the acronym of the scale which belongs to and the number of the item. The stronger relationships are, the thicker the line between nodes are depicted. In A) we show communities at resolution 0.6 and in B) we present communities at resolution 1.

also worth to mention that both approaches reach the best performance using the same layers configuration, namely multi-neuron layers without using bias.

Finally, a short description of each trained model architecture with some training features are drawn in Table 7.1. The elapsed time per training epoch barely suffers when neurons are added, meanwhile the needed number of epochs, until the stop criterion is reached, significantly decreases with multi-neuron configurations. Multi-neuron biased network (CD-DNN_N_B) gives a training error of $9 \cdot 10^{-3}$ while the validation error is $1.8 \cdot 10^{-2}$ when the training process ends. This undesirable behavior is only observed in this con-

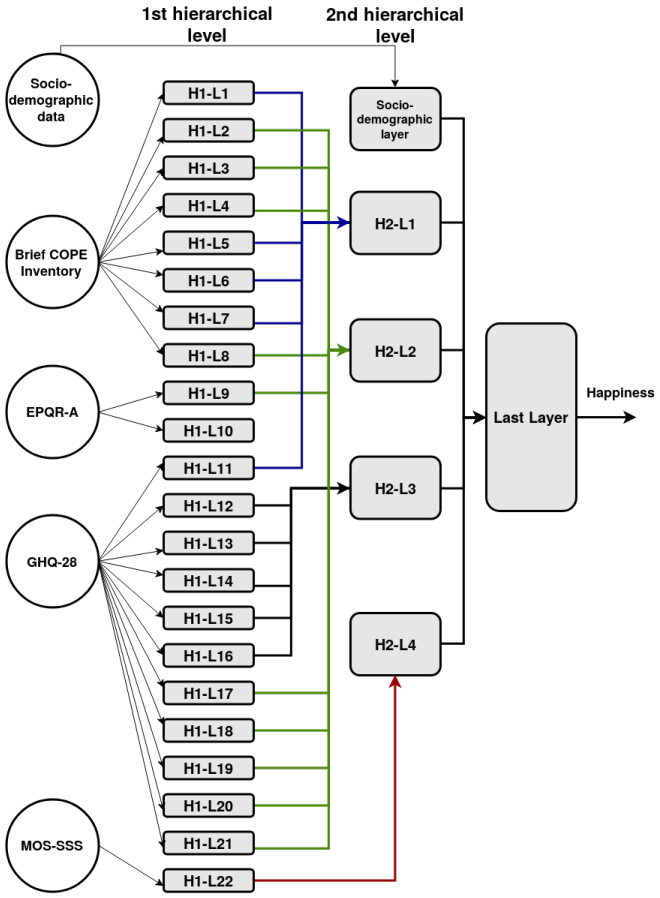


Fig. 7.2: CD-DNN architecture. This architecture is automatically inferred by using the similarity in the responses to the scale items, observed from the responses gathered in the dataset.

figuration. It may imply that this network "learns" so quickly that the stop criterion is not enough to avoid over-fitting problems [233].

Although out of the scope of this work, it is worth to mention that we have noticed latent relationships between the psychometric scales. This can be observed at the inputs of the H_2L_1 and H_2L_2. We observe that the corresponding communities at resolution 1 have grouped altogether layers of different scales in contrast to what can be seen at resolution 0.6 when there were no intersections between scales. It is also highlighted that all the items that belong to the MOS-SSS survey formed one unique community at both resolutions. This suggests that the sub-scales of the MOS-SSS had similar

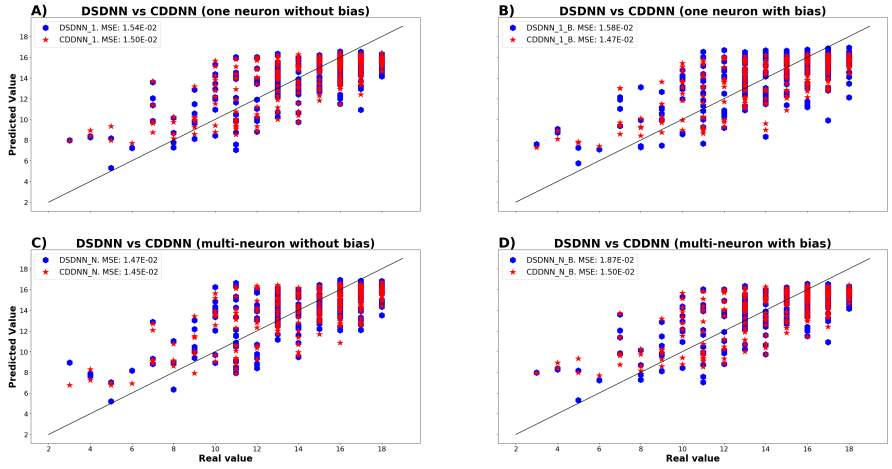


Fig. 7.3: Comparison performance results for DS-DNN and CD-DNN of the tested network configurations. A) presents the comparison of the results of "one-neuron without adding bias" configuration, B) shows the results considering the case of "one-neuron with bias" configuration, C) compares the "multi-neuron without bias" results, and finally D) displays the results for "multi-neuron with bias" configuration.

Table 7.1: The four DNN configurations applied to test the performance of our CD-DNN.

Net	**Bias (T/F)**[1]	**Neurons per layer**	**Epochs**[2]	**Time per epoch**[3]
CDDNN_1	F	1	42	1.32
CDDNN_1_B	T	1	26	1.32
CDDNN_N	F	multi	19	1.34
CDDNN_N_B	T	multi	10	1.35

[1] Use of bias in layers. T = True, F = False.
[2] Number of training epochs until the stop criterion is reached.
[3] Mean elapsed time during one training and validation epoch (in sec).

responses in our dataset and the use of two hierarchical levels may be unnecessary for these items. Finally, 12 of 28 GHQ-28 items were grouped at resolution 1, while these items were grouped into 6 communities at resolution 0.6.

7.5 Discussion

7.5.1 Impact and contributions

The aim of the present study has been the definition of a new methodology to automatically construct the architecture of a deep learning-based model to be used with regression purposes. We focused our efforts on the case Likert-scale datasets. The nature of these measurement artifacts provides a simple way to define similarity relationships between items of different scales.

By definition, the proposed methodology is a contribution in the sense that it automates the construction of the model architecture. Besides, the results showed in Section 7.4 demonstrate that the CD-DNN provides better experimental results on real data when compared with a data-structure architecture drawn using the preconceived sub-scale structure of the scales.

Congruent with our previous work, CD-DNN with unbiased multiple-neuron layers gives the best results. This may indicate that multiple-neuron configurations provide enough adaptability to changes within hierarchical levels, and the learning rate is controllable enough to avoid falling into overfitting problems.

Although the interpretation of the results shown in this paper is out its scope, the presented approach also opens a debate about the possibility of finding latent relationships between items of different psychological scales. Furthermore, the architecture based on these relationships increases the predictive ability of DNN models when compared to those whose architecture is based on the structure of the scales.

Finally, it would be easy to generalize to classification problems by only modifying the behavior of the last layer, which is independent of the community detection based architecture, and the way of codifying the ground truth for supervised learning.

7.5.2 Limitations and future work

Problems based on predictions (either in regression or classification problems) from Likert scales have a very particular field of application. In this sense, the most critical limitation, which is aligned with the most important line for future work, is the capability of extending the automatic construction of the architecture for the DNN for other types of datasets.

If we distinguish between the graph construction and the implementation of the DNN, our efforts should focus on generalizing the graph definition. That is because the DNN is automatically generated when the communities are known, and the communities are automatically detected if the graph is known. So it is enough to find a way to generalize the construction of the graph or to find a set of ways that covered a wide range of data. The use of deep learning in the search of communities could lead to a community detection generalization independent of the data type. It would be necessary

to explore how to mimic the hierarchy for these questionnaires. Although the methodology proposed in the present paper uses agglomerative hierarchical algorithms to match the conceptual structure of the questionnaires, it would also be interesting to try other community detection algorithms where an element could belong to two communities.

The experimental results presented in this work are good for the dataset considered, but the validation of the proposed methodology, considering larger datasets, would also be desirable.

7.6 Conclusion

Deep learning-based models are a trending paradigm in research fields using data science techniques. This technology is increasingly being used for all kinds of tasks and is outperforming results obtained with previous approaches. The main problem of using models based on deep learning is to know the most accurate way of designing them, and many times this process is carried out as trial and error until a quality criterion is reached.

Several studies are based on the analysis of structured data, as is the case with Likert-scales based studies, for which we propose a framework to construct a deep learning model with regression purposes automatically. The methodology consists of the definition theoretic graph quantifying the relationships between different items of the scales, applying a community detection algorithm to infer the architecture for the deep neural network (CDDNN) and the training of the model.

The first experiments using real data have demonstrated better performance when compared to a previous work in which the architecture was drawn by using the conceptual structure of the Likert scales. The natural next step is to get the graph definition abstracted to apply the same methodology to other types of data.

Future steps must explore other approaches for community detection to make the model independent of the data type. Algorithms based on deep learning could deal with community detection generalization.

Acknowledgments

The authors thank the support of the project *Analysis, quality, and variability of medical data* funded by Universitat Politècnica de València. JMGG and JAC acknowledge the support of the H2020 project *CrowdHealth* (Collective Wisdom Driving Public Health Policies - 727560) funded by the European Comission. JMGG acknowledge and to the *InAdvance* project (Patient-Centred Pathways of Early Palliative Care, Supportive Ecosystems and Appraisal Standard - 825750) funded by the European Comission, too.

Conflicts of interest

This work does not have any conflicts of interest.

8 Concluding remarks and recommendations

The human brain is an incredible pattern-matching machine.

Jeff Bezos.

This chapter ends the work carried out in the development of this thesis. A summary of the main concluding remarks is presented below, as well as a set of recommendations for continuing in this line.

8.1 Concluding remarks

Healthcare systems are complex organisms in continuous evolution, and the registries acquired in these scenarios capture this complexity. Besides, the need for much current research requires the combination of data from different healthcare organisms, which promotes the increment of data complexity. It is essential to find methodologies to deal with the challenges that this complexity may raise. This thesis has tackled this problem from two points of view. Firstly, a set of methodologies to prior detection of evidence on differences from distinct sources or time batches was investigated and the way to use this prior information to improve ML-based models' performance. Secondly, it has been demonstrated that letting models be designed using the conceptual structure of data, or even using the observed data, increases the performance without prior knowledge of data distribution. This last has the added value of allowing to extract information of the most influential variables in the prediction tasks.

This thesis has contributed to the fields of Computer Science, Applied Mathematics, Medical Informatics, Medical Imaging, and Biomedical Engineering by the innovative application of trending technologies such as *Data Quality*, *Machine Learning* or *Network Science*. The publication of the de-

veloped work in top-ranked journals and their diffusion in international conferences endorses the community research interest.

The specific concluding remarks of this thesis are listed as follows:

CR1 - Temporal and multisource variability assessment methods are important tools to measure data fitness-to-reuse. Many works have been demonstrated DQ to be a notorious technology that should be applied as the first step when a new database is considered to analyze. Concepts related to a fault acquisition process, such as the lack of complete registries or duplicates ones, are easier to be discovered. In this work, we have focused on the characterization of two inherent-to-data DQ concepts as Temporal and multisource variability are, which could be included inside the *Concordance* and *Timeliness* DQ dimensions. Three key ideas can be highlighted:

- The study conducted in chapter 3 demonstrates that the hospital operations, population changes, and managerial decisions influence the hospital records. These data changes may be monitored by using a TVA method. This method can be useful to measure the impact of managerial decisions on unexpected hospital areas or to know if all the registries would be appropriated for an ML-based task.

  This key idea responds to the research question RQ1, covers the objectives O1 and O3 and derived in the journal publication P1.

- Chapter 4 proposes a semi-automatic methodology to extract the variables influencing a determined classification task. It is tried to classify migraine patients according to the different headache intensities and frequencies. The MVA method endorsed the robustness of the influencing variables in the headache intensity, while it does not provide scientific support on those influencing headache frequency. Random Forest was the ML tool selected to extract the importance of the features. The variables influencing headache intensity were supported by the literature, which signifies that the MVA method could contribute to the credibility of the results when feature extraction uses ML approaches.

  This key idea responds to the research questions RQ1 and RQ2, covers the objectives O1, O2, and O4 and derived in the journal publication P2.

- That the MVA method (and therefore TVA method, which is based on the same technology) could be used to obtain useful information to improve DL model performance is demonstrated in chapter 5. The performance of a DL model was improved by applying a preprocessing step in the histograms of digital mammographies after discovering histogram differences using the MVA method. Here, the histogram can be understood as a PDF, and the MVA method was applied to

each image. Images acquired from different devices showed being far in the simplex built using such a method.

This key idea responds to the research questions RQ1 and RQ2, covers the objectives O1, O2, and O4 and derived in the journal publication P3.

CR2 - The cornerstones for DL emergence have been hardware evolution, an explosion of information sources, and brilliant researchers' imagination. The work conducted in chapter 6 is aligned with such a technique. A methodology to design a DL architecture using the conceptual structure of data (D-SDNN) is proposed. This architecture demonstrated a better performance than state-of-the-art technologies in the psychology field. Besides, the lack of interpretability of DL models is a major drawback in the healthcare environment. It, among other reasons, has lead to the emergence of the "Explainable Artificial Intelligence" (EAI) concept. This concept covers any methodology intended to give insights into the DL-based models working. Besides, two metrics were proposed to measure the influence of each model input in the prediction, which also aligned the EAI topic. Furthermore, the feature importance extraction after the application of the D-SDNN to model a happiness degree predictor provided results concordant with literature.

This concluding remark responds to the research questions RQ3 and RQ4, covers the objectives O2 and O5 and derived in the journal publication P4.

CR3 - It should be noted that the D-SDNN previously proposed cannot deal with the identification of differences among data sources. By definition, it takes the conceptual structure of data to be used in the DL architecture design. Since it uses the observed data, the approach (CD-DNN) described in chapter 7 can somehow take into account the differences in data acquisition. The observed data is used to define a relationship among model inputs. This relationship serves to build an undirected graph where community detection algorithms are applied to obtain the DL architecture. One of the principal advantages of this approach is that it is enough to define a relationship among model inputs for the architecture to be directly built. This framework (CD-DNN) was applied using the same database as used with D-SDNN. The results suggest that this framework can take advantage of the data distribution without any supervision apart from the added-value that the automation signifies. Furthermore, by its definition, the metrics proposed in chapter 6 can still be used in the CD-DNN framework.

This concluding remark responds to the research questions RQ3 and RQ4, covers the objectives O2 and O6 and derived in the journal publication P6 and conferences participation P5 and P7.

8.2 Recommendations

The healthcare systems historically have been a source of large amounts of data. Healthcare data is heterogeneous since it covers socio-demographic, clinical data, or medical imaging to the recent genetic information. Furthermore, such an amount of recorded variables for each patient makes necessary the collection of more and more patients for getting enough dimension to approach the clinical challenges. This supports the need for building large datasets, by combining the registries from several sources and, then, tools to measure the data adequacy or let ML models be designed using the observed data, is becoming crucial.

The research findings and proposed methods discussed in this thesis aim to give some knowledge on that scenario and serve as the starting point for further research. In this sense, the following lines are suggested for future investigations.

R1 - The selected DQ method to assess the multisource and temporal variability is based on the construction of a simplex contained in *Riemannian manifold*. In such a simplex, each point represents a PDF and the distance between two points is the JSD between the PDFs represented by them. The DBScan strategy was applied in that scenario to find clustering patterns that could serve as scientific evidence about the dissimilarity over the found clusters. Despite the appropriateness of using an algorithm based on distance (as DBScan is), there exist several state-of-the-art approaches to deal with this task.
In this regard, we identify this point as a line of future research, where ML approaches can be applied with promising expectations. Supervised learning algorithms such as *K Nearest Neighbors*, *Random Forest*, or *Support Vector Machines*, and Unsupervised learning algorithms, such as *Autoencoders*, can capture subtle clusters that nowadays are not being detected.

R2 - The MVA and TVA methods have demonstrated suiting the task of detecting anomalies in data distributions. These anomalies have also been proved to be effective for designing methodologies that palliate the impact of such for the data reuse. How to join these both paradigms is key to automate the extraction of conclusions with guarantees. Under this assumption, we encourage the use of metrics based on the Information Geometry on which the TVA and MVA methods are founded to train ML models.

R3 - The framework to automatically design DL architectures (CD-DNN) is based on community detection algorithms applied to a weighted-complete graph due to the definition of the relationship between model inputs. A weighted-complete graph is a graph in which each pair of nodes are joint with an associated weight. In this sense, the number of edges is n^2, where n is the number of nodes, and therefore the community detection algorithms have a high computational cost. Since a weighted-complete graph

may be represented by a non-sparse matrix, the community detection strategy that generates the DL architecture could be approached using classical clustering strategies or unsupervised ML algorithms. The current version of the method most probably would need for distributed computation for large datasets.

R4 - Furthermore, the CD-DNN has been validated to be used with a specific data type, Likert-scales. The methodology bases on the application of community detection algorithms at different resolutions, given a graph, which lets to define a hierarchical architecture. This methodology can be available for other data types by defining relationships for the model inputs, which covers all possible data types. In this sense, we encourage to define relationships between model inputs with a wide range of data covered or to explore relationships that could be applied to any data types.
It would also be desirable to validate the performance of the automatic deep neural network design in larger datasets that covered multiple sources.

R5 - Breast cancer is one of the most frequent cancers in women. The screening programs have emerged with the purpose of early diagnostic. The percentage of dense tissue over the breast is demonstrated to be an important biomarker of the risk of disease development. Although these programs are intended to discover malignancies in breasts, obtaining an objective way to measure this marker is crucial. The differences inter and intra reader, the differences in the acquisition process of the devices, or if the mammograms are stored processed or not, have a high impact on the standardization of the measurement. Besides, it could serve, for example, to generate decision-making aid systems for the establishment of women follow-up periodicity.
This thesis was not intended to the breast percent density reading standardization, but the results obtained in Chapter 5 suggest that it is an important field.

References

[1] R.-Y. Wang and D.-M. Strong, "Beyond accuracy: What data quality means to data consumers," *J. Manage. Inform. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[2] I.-G. Todoran, L. Lecornu, A. Khenchaf, and J.-M.-L. Caillec, "A methodology to evaluate important dimensions of information quality in systems," *Journal of Data and Information Quality (JDIQ)*, vol. 6, no. 2-3, pp. 1–23, 2015.

[3] M.-G. Kahn, M.-A. Raebel, J.-M. Glanz, K. Riedlinger, and J.-F. Steiner, "A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research," *Med. Care*, vol. 50, 2012.

[4] C. Sáez, M. Robles, and J.-M. García-Gómez, "Stability metrics for multi-source biomedical data based on simplicial projections from probability distribution distances," *Statistical methods in medical research*, vol. 26, no. 1, pp. 312–336, 2017.

[5] C. Sáez, P.-P. Rodrigues, J. Gama, M. Robles, and J.-M. García-Gómez, "Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality," *Data Min. Knowl. Disc.*, vol. 29, no. 4, pp. 950–975, 2015.

[6] N.-S. Rajan, R. Gouripeddi, P. Mo, R.-K. Madsen, and J.-C. Facelli, "Towards a content agnostic computable knowledge repository for data quality assessment," *Comput. Meth. Prog. Bio.*, vol. 177, pp. 193–201, 2019.

[7] T. Callahan, J. Barnard, L. Helmkamp, J. Maertens, and M. Kahn, "Reporting data quality assessment results: identifying individual and organizational barriers and solutions," *eGEMs*, vol. 5, no. 1, 2017.

[8] M.-G. Kahn, J.-S. Brown, A.-T. Chun, B.-N. Davidson, D. Meeker, P.-B. Ryan, L.-M. Schilling, N.-G. Weiskopf, A.-E. Williams, and M.-N. Zozus, "Transparent reporting of data quality in distributed data networks," *eGEMs*, vol. 3, no. 1, 2015.

[9] N.-G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *J. Am. Med. Inform. Assn.*, vol. 20, no. 1, pp. 144–151, 2013.

[10] O. Almutiry, G. Wills, and R. Crowder, "A dimension-oriented taxonomy of data quality problems in electronic health records," *IADIS International Journal on WWW/Internet*, vol. 13, no. 2, pp. 98–114, 2016.

[11] S.-G. Johnson, S. Speedie, G. Simon, V. Kumar, and B.-L. Westra, "A Data Quality Ontology for the Secondary Use of EHR Data.," *AMIA Annu. Symp. Proc.*, vol. 2015, pp. 1937–46, 2015.

[12] M.-G. Kahn, T.-J. Callahan, J. Barnard, A.-E. Bauck, J. Brown, B.-N. Davidson, H. Estiri, C. Goerg, E. Holve, S.-G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T.-C. Ong, P. Ryan, N. Shang, N.-G. Weiskopf, C. Weng, M.-N. Zozus, and L. Schilling, "A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data," *eGEMs*, vol. 4, p. 18, 9 2016.

[13] A.-M. Turing, "Computing Machinery and Intelligence," in *Parsing the Turing Test*, pp. 23–65, Dordrecht: Springer Netherlands, 2009.

[14] R. J. Wieringa, *Design science methodology for information systems and software engineering*. Springer, 2014.

[15] C.-E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[16] S.-I. Amari and H. Nagaoka, *Methods of information geometry*, vol. 191. American Mathematical Soc., 2007.

[17] E. Parzen, "On estimation of a probability density function and mode," *The Annals of Mathematical Statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.

[18] C. Sáez, *Probabilistic methods for multi-source and temporal biomedical data quality assessment*. PhD thesis, Editorial Universitat Politècnica de València, 2016.

[19] H.-R. Parks and D.-C. Wills, "An elementary calculation of the dihedral angle of the regular n-simplex," *Am. Math. Mon.*, vol. 109, no. 8, pp. 756–758, 2002.

[20] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8599–8603, IEEE, 5 2013.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, 6 2016.

[22] M. Moravčík, M. Schmid, N. Burch, V. Lisý, D. Morrill, N. Bard, T. Davis, K. Waugh, M. Johanson, and M. Bowling, "DeepStack: Expert-level artificial intelligence in heads-up no-limit poker," *Science*, vol. 356, pp. 508–513, 5 2017.

[23] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," *Empir. Soft. Eng.*, vol. 19, pp. 465–500, 8 2017.

[24] J. D, E. H, and Y. S, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res*, vol. 12, no. 61, pp. 2121–2159, 2011.

[25] D.-P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] S. Dutta, S. Jha, S. Sankaranarayanan, and A. Tiwari, "Output range analysis for deep feedforward neural networks," in *NASA Formal Methods Symposium*, pp. 121–138, Springer, 2018.

[27] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *11th Annual Conference of the International Speech Communication Association (ISCA)*, 2010.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[29] A. Krizhevsky, I. Sutskever, and G.-E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Adv. Neural Inf. Process. Syst.*, pp. 1097–1105, 2012.

[30] L. Euler, "The seven bridges of königsberg," *The world of mathematics*, vol. 1, pp. 573–580, 1956.

[31] P. Singh, J. Guo, P. Bhandary, E. Wurtele, and K. Bassler, "A novel community detection method improves detection of functional gene modules in big gene expression data.," *Bull. Am. Phys. Soc.*, 2020.

[32] S. Javed, A. Mahmood, M.-M. Fraz, N.-A. Koohbanani, K. Benes, Y.-W. Tsang, K. Hewitt, D. Epstein, D. Snead, and N. Rajpoot, "Cellular community detection for tissue phenotyping in colorectal cancer histology images," *Med. Image Anal*, p. 101696, 2020.

[33] Z. Yang, R. Algesheimer, and C.-J. Tessone, "A comparative analysis of community detection algorithms on artificial networks," *Sci. Rep.*, vol. 6, p. 30750, 2016.

[34] S. Fortunato, "Community detection in graphs," *Phys. Rep.*, vol. 486, no. 3-5, pp. 75–174, 2010.

[35] B.-W. K. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *The Bell System Technical Journal*, vol. 49, no. 2, pp. 291–307, 1970.

[36] M.-E.-J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, p. 026113, 2004.

[37] V.-D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech. Theory Exp*, vol. 2008, no. 10, p. P10008, 2008.

[38] R.-S. Aguilar-Saven, "Business process modelling: Review and framework," *Int. J. Prod. Econ.*, vol. 90, no. 2, pp. 129–149, 2004.

[39] M. Poulymenopoulou, F. Malamateniou, and G. Vassilacopoulos, "Specifying workflow process requirements for an emergency medical service," *J. Med. Syst.*, vol. 27, no. 4, pp. 325–335, 2003.

[40] P. Dadam, M. Reichert, and K. Kuhn, "Clinical workflowsâ€"the killer application for process-oriented information systems?," in *Proceedings of the 4th International Conference on Business Information Systems*, pp. 36–59, Springer, 2000.

[41] R. Lenz and M. Reichert, "It support for healthcare processes–premises, challenges, perspectives," *Data Knowl. Eng.*, vol. 61, no. 1, pp. 39–58, 2007.

[42] K. Anyanwu, A. Sheth, J. Cardoso, J. Miller, K. Kochut, *et al.*, "Healthcare enterprise process development and integration," *J. Res. Pract. Inf.Technol.*, vol. 35, no. 2, p. 83, 2003.

[43] A. Rebuge and D. R. Ferreira, "Business process analysis in healthcare environments: A methodology based on process mining," *Inf. Syst.*, vol. 37, no. 2, pp. 99–116, 2012.

[44] E. A. E. H. Amor and S.-A. Ghannouchi, "Applying data mining techniques to discover kpis relationships in business process context," in *18th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT)*, pp. 230–237, IEEE, 2017.

[45] Y.-C. Chou, B.-Y. Chen, Y.-Y. Tang, Z.-J. Qiu, M.-F. Wu, S.-C. Wang, H.-S. Lin, and W.-C. Chuang, "Prescription-filling process reengineering of an outpatient pharmacy," *J. Med. Syst.*, vol. 36, no. 2, pp. 893–902, 2012.

[46] J.-D. Leu and Y.-T. Huang, "An application of business process method to the clinical efficiency of hospital," *J. Med. Syst.*, vol. 35, no. 3, pp. 409–421, 2011.

[47] K. Gand, "Investigating on requirements for business model representations: The case of information technology in healthcare," in *19th IEEE Conference on Business Informatics (CBI)*, vol. 1, pp. 471–480, IEEE, 2017.

[48] G.-S.-A. Ferreira, U.-R. Silva, A.-L. Costa, and S.-D. de Dallavalle Pádua, "The promotion of bpm and lean in the health sector: main results," *Bus. Process Manag. J.*, 2018.

[49] A.-M. Jabour, *Cancer reporting: timeliness analysis and process reengineering*. PhD thesis, 2015.

[50] M. Hewitt and J.-V. Simone, *Enhancing data systems to improve the quality of cancer care*. Natl. Academy Pr., 2000.

[51] C. Sáez, J. Martínez-Miranda, M. Robles, and J.-M. García-Gómez, "Organizing data quality assessment of shifting biomedical data.," *Studies in Health Record Data*, vol. 180, pp. 721–725, 2012.

[52] C. Sáez, M. Robles, and J.-M. Garcia-Gomez, "Comparative study of probability distribution distances to define a metric for the stability of multi-source biomedical research data," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 3226–3229, IEEE, 2013.

[53] "Universitari i politècnic la fe hospital research ethics committee.."

[54] S. Rose, "International ethical guidelines for epidemiological studies: by the council for international organizations of medical sciences (cioms)," 2009.

[55] M.-E. Charlson, P. Pompei, K.-L. Ales, and C.-R. MacKenzie, "A new method of classifying prognostic comorbidity in longitudinal studies: development and validation," *Chronic Dis.*, vol. 40, no. 5, pp. 373–383, 1987.

[56] S. Schneeweiss, P.-S. Wang, J. Avorn, and R.-J. Glynn, "Improved comorbidity adjustment for predicting mortality in medicare populations," *Health Serv.Res.*, vol. 38, no. 4, pp. 1103–1120, 2003.

[57] H. Quan, V. Sundararajan, P. Halfon, A. Fong, B. Burnand, J.-C. Luthi, L.-D. Saunders, C.-A. Beck, T.-E. Feasby, and W.-A. Ghali, "Coding algorithms for defining comorbidities in icd-9-cm and icd-10 administrative data," *Med. Care*, pp. 1130–1139, 2005.

[58] C. Sáez, O. Zurriaga, J. Pérez-Panadés, I. Melchor, M. Robles, and J.-M. García-Gómez, "Applying probabilistic temporal and multisite data quality control methods to a public health mortality registry in spain: a systematic approach to quality control of repositories," *J. Am. Med. InformaticsAssoc.*, vol. 23, no. 6, pp. 1085–1095, 2016.

[59] I. Csiszár and P.-C. Shields, *Information theory and statistics: A tutorial*. Found. Trends Commun. Inf. Theory., 2004.

[60] B.-A. Turlach, "Bandwidth selection in kernel density estimation: A review," in *CORE and Institut de Statistique*, Citeseer, Universitè Catholique de Louvain, Louvain-la-Neuve, 1993.

[61] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Trans. Inf.Theory.*, vol. 37, no. 1, pp. 145–151, 1991.

[62] T.-M. Cover and J.-A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[63] I.-T. Jolliffe, "Principal components in regression analysis," in *Principal component analysis*, pp. 129–155, Springer, 1986.

[64] M.-L. Davison and S.-G. Sireci, "Multidimensional scaling," in *Handbook of applied multivariate statistics and mathematical modeling*, pp. 323–352, Elsevier, 2000.

[65] U. Brandes and C. Pich, "Eigensolver methods for progressive multidimensional scaling of large data," in *International Symposium on Graph Drawing*, pp. 42–53, Springer, 2006.

[66] M. Daszykowski and B. Walczak, "Density-based clustering methods," *Comprehensive Chemoetrics*, 2009.

[67] S.-T. Liaw, A. Rahimi, P. Ray, J. Taggart, S. Dennis, S. de Lusignan, B. Jalaludin, A.-E.-T. Yeo, and A. Talaei-Khoei, "Towards an ontology for data quality in integrated chronic disease management: a realist review of the literature," *Int. J. Med. Inform.*, vol. 82, no. 1, pp. 10–24, 2013.

[68] D.-M. Parkin and F. Bray, "Evaluation of data quality in the cancer registry: principles and methods part ii. completeness," *Eur. J Cancer.*, vol. 45, no. 5, pp. 756–764, 2009.

[69] C. Fernandez-Llatas, G. Ibanez-Sanchez, A. Celda, J. Mandingorra, L. Aparici-Tortajada, A. Martinez-Millana, J. Munoz-Gama, M. Sepúlveda, E. Rojas, V. Gálvez, *et al.*, "Analyzing medical emergency processes with process mining: the stroke case," in *International Conference on Business Process Management*, pp. 214–225, Springer, 2018.

[70] C. Fernandez-Llatas, A. Lizondo, E. Monton, J.-M. Benedi, and V. Traver, "Process mining methodology for health process tracking using real-time indoor location systems," *Sensors*, vol. 15, no. 12, pp. 29821–29840, 2015.

[71] B.-F. Van Dongen, A.-K.-A. de Medeiros, H.-M.-W. Verbeek, A.-J.-M.-M. Weijters, and W.-M.-P. van Der Aalst, "The prom framework: A new era in process mining tool support," in *International Conference on Application and Theory of Petri Nets*, pp. 444–454, Springer, Springer, 2005.

[72] W. Van der Aalst, T. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, 2004.

[73] A.-K.-A. De Medeiros, A.-J.-M.-M. Weijters, and W.-M.-P. Van der Aalst, "Genetic process mining: a basic approach and its challenges," in *International Conference on Business Process Management*, pp. 203–215, Springer, Springer, 2005.

[74] A.-J.-M.-M. Weijters, W.-M.-P. van Der Aalst, and A.-K.-A. De Medeiros, "Process mining with the heuristics miner-algorithm," *Technische Universiteit Eindhoven, Tech. Rep. WP*, vol. 166, pp. 1–34, 2006.

[75] J. Barjis, A. Verbraeck, S.-J. Shim, and A. Kumar, "Simulation for emergency care process reengineering in hospitals," *Bus. Process Manag. J.*, 2010.

[76] G. Svolba and P. Bauer, "Statistical quality control in clinical trials," *Control Clin. Trials.*, vol. 20, no. 6, pp. 519–530, 1999.

[77] F. Bray and D.-M. Parkin, "Evaluation of data quality in the cancer registry: principles and methods. part i: comparability, validity and timeliness," *Eur. J. Cancer.*, vol. 45, no. 5, pp. 747–755, 2009.

[78] M.-G. Kahn, M.-A. Raebel, J.-M. Glanz, K. Riedlinger, and J.-F. Steiner, "A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research," *Med. Care.*, vol. 50 (SUPPL. 1) S21–9, 2012.

[79] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–52, 2009.

[80] B. Heinrich, M. Klier, and M. Kaiser, "A procedure to develop metrics for currency and its application in crm," *J. Data Inf. Qual.*, vol. 1, no. 1, pp. 1–28, 2009.

[81] G. Sirgo, F. Esteban, J. Gómez, G. Moreno, A. Rodríguez, L. Blanch, J.-J. Guardiola, R. Gracia, L. De Haro, and M. Bodí, "Validation of the icu-dama tool for automatically extracting variables for minimum dataset and quality indicators: The importance of data quality assessment," *Int. J. Med. Inform.*, vol. 112, pp. 166–172, 2018.

[82] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach.Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[83] G.-E. Hinton and R.-R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science.*, vol. 313, no. 5786, pp. 504–507, 2006.

[84] K.-Q. Weinberger, F. Sha, and L.-K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Inf. Conf. Mach. Learn.*, pp. 106–113, 2004.

[85] G. Fanjiang, J.-H. Grossman, W.-D. Compton, P.-P. Reid, *et al.*, *Building a better delivery system: a new engineering/health care partnership*. National Academies Press., 2005.

[86] L.-T. Kohn, J. Corrigan, M.-S. Donaldson, *et al.*, *To err is human: building a safer health system*, vol. 6. National Academy Press., 2000.

[87] Y.-W. Woldeamanuel and R.-P. Cowan, "Migraine affects 1 in 10 people worldwide featuring recent rise: a systematic review and meta-analysis of community-based studies involving 6 million participants," *J. Neurol. Sci.*, vol. 372, pp. 307–315, 2017.

[88] T. Vos, A.-A. Abajobir, K.-H. Abate, C. Abbafati, K.-M. Abbas, F. Abd-Allah, R.-S. Abdulkader, A.-M. Abdulle, T.-A. Abebo, S.-F. Abera, *et al.*, "Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the global burden of disease study 2016," *Lancet*, vol. 390, no. 10100, pp. 1211–1259, 2017.

[89] A. Raggi and M. Leonardi, "Burden and cost of neurological diseases: a e uropean n orth–s outh comparison," *Acta Neurol. Scand.*, vol. 132, no. 1, pp. 16–22, 2015.

[90] M. de Tommaso and V. Sciruicchio, "Migraine and central sensitization: clinical features, main comorbidities and therapeutic perspectives," *Curr.Rheumatol. Rev.*, vol. 12, no. 2, pp. 113–126, 2016.

[91] K.-Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," *Lancet Oncol.*, vol. 20, no. 5, pp. e262–e273, 2019.

[92] F.-J. Pérez-Benito, P. Villacampa-Fernández, J.-A. Conejero, J.-M. García-Gómez, and E. Navarro-Pardo, "A happiness degree predictor using the conceptual data structure for deep learning architectures," *Comput. Methods Programs Biomed*, vol. 168, pp. 59–68, 2019.

[93] E. Navarro-Pardo, L. González-Pozo, P. Villacampa-Fernández, and J.-A. Conejero, "Benefits of a dance group intervention on institutionalized elder people: a bayesian network approach," *Applied Mathematics and Nonlinear Sciences*, vol. 3, no. 2, pp. 503–512, 2018.

[94] A.-P. Reimer, N.-K. Schiltz, V.-P. Ho, E.-A. Madigan, and S.-M. Koroukian, "Applying supervised machine learning to identify which patient characteristics identify the highest rates of mortality post-interhospital transfer," *Biomed.Inform. Insights*, vol. 11, p. 1178222619835548, 2019.

[95] Y. Garcia-Chimeno, B. Garcia-Zapirain, M. Gomez-Beldarrain, B. Fernandez-Ruanova, and J.-C. Garcia-Monco, "Automatic migraine classification via feature selection committee and machine learning techniques over imaging and questionnaire data," *BMC Med. Inform. Decis. Mak.*, vol. 17, no. 1, p. 38, 2017.

[96] J. Lötsch and A. Ultsch, "Machine learning in pain research," *Pain*, vol. 159, no. 4, p. 623, 2018.

[97] "Ichd-iii headache classification subcommitte of the international headachesociety: The international classification of headache disorders, 3 edition.," *Cephalagia*, vol. 38, pp. 1–211, 2018.

[98] Z. Liang, O. Galea, L. Thomas, G. Jull, and J. Treleaven, "Cervical musculoskeletal impairments in migraine and tension type headache: A systematic review and meta-analysis," *Musculoskelet. Sci. Pract.*, 2019.

[99] D. Phillip, A.-C. Lyngberg, and R. Jensen, "Assessment of headache diagnosis. a comparative population study of a clinical interview with a diagnostic headache diary," *Cephalalgia*, vol. 27, no. 1, pp. 1–8, 2007.

[100] A.-S. Zigmond and R.-P. Snaith, "The hospital anxiety and depression scale," *Acta Psychiatr. Scand.*, vol. 67, no. 6, pp. 361–370, 1983.

[101] K.-D. Juang, S.-J. Wang, C.-H. Lin, and J.-L. Fuh, "Use of the hospital anxiety and depression scale as a screening tool for patients with headache.," *ZhonghuaYi Xue Za Zhi (Taipei)*, vol. 62, no. 11, pp. 749–755, 1999.

[102] C.-D. Spielberger, "State-trait anxiety inventory: A comprehensive bibliography," *Palo Alto, CA: Consulting Psychologists Press*, 1989.

[103] L.-L.-B. Barnes, D. Harp, and W.-S. Jung, "Reliability generalization of scores on the spielberger state-trait anxiety inventory," *Educ. Psychol. Meas.*, vol. 62, no. 4, pp. 603–618, 2002.

[104] W.-F. Stewart, R.-B. Lipton, K.-B. Kolodner, J. Sawyer, C. Lee, and J.-N. Liberman, "Validity of the migraine disability assessment (midas) score in comparison to a diary-based measure in a population sample of migraine sufferers," *Pain*, vol. 88, no. 1, pp. 41–52, 2000.

[105] M. Palacios-Ceña, L. Lima Florencio, G. Natália Ferracini, J. Barón, Á.-L. Guerrero, C. Ordás-Bandera, L. Arendt-Nielsen, and C. Fernández-de Las-Peñas, "Women with chronic and episodic migraine exhibit sim-

ilar widespread pressure pain sensitivity," *Pain Med.*, vol. 17, no. 11, pp. 2127–2133, 2016.

[106] K. Luedtke, W. Boissonnault, N. Caspersen, R. Castien, A. Chaibi, D. Falla, C. Fernandez-de-las Penas, T. Hall, J.-R. Hirsvang, T. Horre, *et al.*, "International consensus on the most useful physical examination tests used by physiotherapists for patients with headache: A delphi study," *Man. Ther.*, vol. 23, pp. 17–24, 2016.

[107] A.-I.-S. Oliveira-Souza, L.-L. Florencio, G.-F. Carvalho, C. Fernández-De-Las-Peñas, F. Dach, and D. Bevilaqua-Grossi, "Reduced flexion rotation test in women with chronic and episodic migraine," *Braz. J. Phys.Ther.*, vol. 23, no. 5, pp. 387–394, 2019.

[108] D. Bevilaqua-Grossi, K. S. Pegoretti, M.-C. Goncalves, J.-G. Speciali, C.-A. Bordini, and M.-E. Bigal, "Cervical mobility in women with migraine," *Headache: The Journal of Head and Face Pain*, vol. 49, no. 5, pp. 726–731, 2009.

[109] G.-N. Ferracini, L.-L. Florencio, F. Dach, D.-G. Bevilaqua, M. Palacios-Cena, C. Ordas-Bandera, T.-C. Chaves, J.-G. Speciali, and C. Fernández-de Las-Peñas, "Musculoskeletal disorders of the upper cervical spine in women with episodic or chronic migraine.," *Eur. J.Phys. Rehabil. Med.*, vol. 53, no. 3, pp. 342–350, 2017.

[110] C. Fernández-de Las-Peñas, M.-L. Cuadrado, and J.-A. Pareja, "Myofascial trigger points, neck mobility and forward head posture in unilateral migraine," *Cephalalgia*, vol. 26, no. 9, pp. 1061–1070, 2006.

[111] K. Luedtke and A. May, "Stratifying migraine patients based on dynamic pain provocation over the upper cervical spine," *J. Headache Pain*, vol. 18, no. 1, p. 97, 2017.

[112] F.-J. Pérez-Benito, C. Sáez, J.-A. Conejero, S. Tortajada, B. Valdivieso, and J.-M. García-Gómez, "Temporal variability analysis reveals biases in electronic health records due to hospital process reengineering interventions over seven years," *PLoS One*, vol. 14, no. 8, p. e0220369, 2019.

[113] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: a critical evaluation," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 3, p. 74, 2016.

[114] 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.

[115] B.-N.-I. Eskelson, H. Temesgen, V. Lemay, T.-M. Barrett, N.-L. Crookston, and A.-T. Hudak, "The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases," *Scand. J. For. Res.*, vol. 24, no. 3, pp. 235–246, 2009.

[116] J.-M. Jerez, I. Molina, P.-J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco, "Missing data imputation using statistical

and machine learning methods in a real breast cancer problem," *Artif. Intell. Med.*, vol. 50, no. 2, pp. 105–115, 2010.

[117] N.-B. Heidenreich, A. Schindler, and S. Sperlich, "Bandwidth selection for kernel density estimation: a review of fully automatic selectors," *AStA Advances in Statistical Analysis*, vol. 97, no. 4, pp. 403–433, 2013.

[118] F.-W. Young and R.-M. Hamer, *Multidimensional scaling: History, theory, and applications*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1987.

[119] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[120] T.-M. Oshiro, P.-S. Perez, and J.-A. Baranauskas, "How many trees in a random forest?," in *Springer 8th International Conference on Machine Learning and Data Mining in Pattern Recognition*, pp. 154–168, Springer, 2012.

[121] B. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams, and H. Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, no. 13, pp. 1636–1643, 2003.

[122] M. Liu, M. Wang, J. Wang, and D. Li, "Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and chinese vinegar," *Sensors and Actuators B. Chem.*, vol. 177, pp. 970–980, 2013.

[123] T.-P. Vital, M.-M. Krishna, G.-V.-L. Narayana, P. Suneel, and P. Ramarao, "Empirical analysis on cancer dataset with machine learning algorithms," in *Adv. Intelligent Systems Computing*, vol. 758, pp. 789–801, 2019.

[124] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res*, vol. 12, pp. 2825–2830, 2011.

[125] J.-S. Chou, C.-K. Chiu, M. Farfoura, and I. Al-Taharwa, "Optimizing the prediction accuracy of concrete compressive strength based on a comparison of data-mining techniques," *J. Computing Civil Engineering*, vol. 25, no. 3, pp. 242–253, 2011.

[126] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *Lecture Notes in Computer Science*, vol. 3048, pp. 345–359, Springer, 2005.

[127] M.-M. Bragatto, D. Bevilaqua-Grossi, M.-T. Benatto, S.-S. Lodovichi, C.-F. Pinheiro, G.-F. Carvalho, F. Dach, C. Fernández-de-las Peñas, and L.-L. Florencio, "Is the presence of neck pain associated with more severe clinical presentation in patients with migraine? a cross-sectional study," *Cephalalgia*, vol. 39, no. 12, pp. 1500–1508, 2019.

[128] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognit. Lett*, vol. 15, no. 11, pp. 1119–1125, 1994.

[129] C.-K. Kuhl, "The changing world of breast cancer: a radiologist's perspective," *Invest. Radiol.*, vol. 50, no. 9, p. 615, 2015.

[130] N.-F. Boyd, J.-M. Rommens, K. Vogt, V. Lee, J.-L. Hopper, M.-J. Yaffe, and A.-D. Paterson, "Mammographic breast density as an intermediate phenotype for breast cancer," *The Lancet Oncology*, vol. 6, no. 10, pp. 798–808, 2005.

[131] V. Assi, J. Warwick, J. Cuzick, and S.-W. Duffy, "Clinical and epidemiological issues in mammographic density," *Nature Reviews Clinical Oncology*, vol. 9, no. 1, p. 33, 2012.

[132] C.-J. D'Orsi, E.-A. Sickles, E.-B. Mendelson, and E.-A. Morris, *ACR BI-RADS® Atlas, Breast Imaging Reporting and Data System*. Reston, VA, American College of Radiology, 2013.

[133] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Pérez, E.-R.-E. Denton, and R. Zwiggelaar, "A novel breast tissue density classification methodology," *IEEE T. Inf. Technol. B.*, vol. 12, no. 1, pp. 55–65, 2008.

[134] F.-J. Pérez-Benito, F. Signol, J.-C. Perez-Cortes, M. Pollán, B. Pérez-Gómez, D. Salas-Trejo, M. Casals, I. Martínez, and R. LLobet, "Global parenchymal texture features based on histograms of oriented gradients improve cancer development risk estimation from healthy breasts," *Comput. Meth. Prog. Bio.*, vol. 177, pp. 123–132, 2019.

[135] S. Ciatto, N. Houssami, A. Apruzzese, E. Bassetti, B. Brancato, F. Carozzi, S. Catarzi, M.-P. Lamberini, G. Marcelli, R. Pellizzoni, *et al.*, "Categorizing breast mammographic density: intra-and interobserver reproducibility of bi-rads density categories," *The Breast*, vol. 14, no. 4, pp. 269–275, 2005.

[136] P. Skaane, "Studies comparing screen-film mammography and full-field digital mammography in breast cancer screening: updated review," *Acta Radiologica*, vol. 50, no. 1, pp. 3–14, 2009.

[137] D. van der Waal, G.-J. den Heeten, R.-M. Pijnappel, K.-H. Schuur, J.-M.-H. Timmers, A.-L.-M. Verbeek, and M.-J.-M. Broeders, "Comparing visually assessed bi-rads breast density and automated volumetric breast density software: a cross-sectional study in a breast cancer screening setting," *PLoS One*, vol. 10, no. 9, p. e0136667, 2015.

[138] S.-H. Kim, E.-H. Lee, J.-K. Jun, Y.-M. Kim, Y.-W. Chang, J.-H. Lee, H.-W. Kim, E. J. Choi, *et al.*, "Interpretive performance and interobserver agreement on digital mammography test sets," *Korean journal of radiology*, vol. 20, no. 2, pp. 218–224, 2019.

[139] K.-J. Geras, R.-M. Mann, and L. Moy, "Artificial intelligence for mammography and digital breast tomosynthesis: Current concepts and future perspectives," *Radiology*, p. 182627, 2019.

[140] R. Miotto, F. Wang, S. Wang, X. Jiang, and J.-T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2017.

[141] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[142] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Proc. Mag.*, vol. 29, 2012.

[143] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recogn. Lett.*, vol. 119, pp. 3–11, 2019.

[144] M. Helmstaedter, K.-L. Briggman, S.-C. Turaga, V. Jain, H.-S. Seung, and W. Denk, "Connectomic reconstruction of the inner plexiform layer in the mouse retina," *Nature*, vol. 500, no. 7461, p. 168, 2013.

[145] K. Lee, N. Turner, T. Macrina, J. Wu, R. Lu, and H.-S. Seung, "Convolutional nets for reconstructing neural circuits from brain images acquired by serial section electron microscopy," *Curr. Opin. Neurobiol.*, vol. 55, pp. 188–198, 2019.

[146] M.-K.-K. Leung, H.-Y. Xiong, L.-J. Lee, and B.-J. Frey, "Deep learning of the tissue-regulated splicing code," *Bioinformatics*, vol. 30, no. 12, pp. i121–i129, 2014.

[147] J. Zhou, C.-Y. Park, C.-L. Theesfeld, A.-K. Wong, Y. Yuan, C. Scheckel, J.-J. Fak, J. Funk, K. Yao, Y. Tajima, *et al.*, "Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk," *Nat. Genet.*, vol. 51, no. 6, p. 973, 2019.

[148] M. Kallenberg, K. Petersen, M. Nielsen, A.-Y. Ng, P. Diao, C. Igel, C.-M. Vachon, K. Holland, R.-R. Winkel, N. Karssemeijer, *et al.*, "Unsupervised deep learning applied to breast density segmentation and mammographic risk scoring," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1322–1331, 2016.

[149] S.-K. Zhou, H. Greenspan, and D. Shen, *Deep learning for medical image analysis*. Academic Press, 2017.

[150] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[151] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. CVPR. IEEE*, pp. 648–656, 2015.

[152] Y. Taigman, M. Yang, M.-A. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. CVPR. IEEE*, pp. 1701–1708, 2014.

[153] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *arXiv preprint arXiv:1312.6229*, 2013.

[154] L.-R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[155] M. Pollán, R. Llobet, J. Miranda-García, J. Antón, M. Casals, I. Martínez, C. Palop, F. Ruiz-Perales, C. Sánchez-Contador, C. Vidal, *et al.*, "Validation of dm-scan, a computer-assisted tool to assess mammographic density in full-field digital mammograms," *Springerplus*, vol. 2, no. 1, p. 242, 2013.

[156] R. Llobet, M. Pollán, J. Antón, J. Miranda-García, M. Casals, I. Martínez, F. Ruiz-Perales, B. Pérez-Gómez, D. Salas-Trejo, and J.-C. Perez-Cortes, "Semi-automated and fully automated mammographic density measurement and breast cancer risk prediction," *Comput Methods Programs Biomed*, vol. 116, no. 2, pp. 105–115, 2014.

[157] L. He, X. Ren, Q. Gao, X. Zhao, B. Yao, and Y. Chao, "The connected-component labeling problem: A review of state-of-the-art algorithms," *Pattern Recognit.*, vol. 70, pp. 25–43, 2017.

[158] K. Wu, E. Otoo, and K. Suzuki, "Optimizing two-pass connected-component labeling algorithms," *Pattern Anal. Appl.*, vol. 12, no. 2, pp. 117–135, 2009.

[159] R. Lakshmanan, V. Thomas, S.-M. Jacob, P. Thara, *et al.*, "Pectoral muscle boundary detection in mammograms using homogeneous contours," in *2015 Fifth International Conference on Advances in Computing and Communications (ICACC)*, pp. 354–357, IEEE, 2015.

[160] R. Shen, K. Yan, F. Xiao, J. Chang, C. Jiang, and K. Zhou, "Automatic pectoral muscle region segmentation in mammograms using genetic algorithm and morphological selection," *J. Digit. Imaging*, vol. 31, no. 5, pp. 680–691, 2018.

[161] K. Yin, S. Yan, C. Song, and B. Zheng, "A robust method for segmenting pectoral muscle in mediolateral oblique (mlo) mammograms," *Int. J. Camput. Ass. Rad.*, vol. 14, no. 2, pp. 237–248, 2019.

[162] V. Shinde and B.-T. Rao, "Novel approach to segment the pectoral muscle in the mammograms," in *Cognitive Informatics and Soft Computing*, pp. 227–237, Springer, 2019.

[163] J.-J. James, "The current status of digital mammography," *Clin. Radiol.*, vol. 59, no. 1, pp. 1–10, 2004.

[164] W. He, S. Harvey, A. Juette, E.-R.-E. Denton, and R. Zwiggelaar, "Mammographic segmentation and density classification: a fractal inspired approach," in *International Workshop on Breast Imaging*, pp. 359–366, Springer, 2016.

[165] A.-K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recogn. Lett.*, vol. 31, no. 8, pp. 651–666, 2010.

[166] J. Lee and R.-M. Nishikawa, "Automated mammographic breast density estimation using a fully convolutional network," *Med. Phys.*, vol. 45, no. 3, pp. 1178–1190, 2018.

[167] T. Buelow, H.-S. Heese, R. Grewer, D. Kutra, and R. Wiemker, "Inter- and intra-observer variations in the delineation of lesions in mammograms," in *Medical Imaging 2015: Image Perception, Observer Performance, and Technology Assessment*, vol. 9416, p. 941605, International Society for Optics and Photonics, 2015.

[168] W. Alakwaa, M. Nassef, and A. Badr, "Lung cancer detection and classification with 3d convolutional neural network (3d-cnn)," *Lung Cancer*, vol. 8, no. 8, p. 409, 2017.

[169] N. Wu, K.-J. Geras, Y. Shen, J. Su, S.-G. Kim, E. Kim, S. Wolfson, L. Moy, and K. Cho, "Breast density classification with deep convolutional neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6682–6686, IEEE, 2018.

[170] C.-D. Lehman, A. Yala, T. Schuster, B. Dontchos, M. Bahl, K. Swanson, and R. Barzilay, "Mammographic breast density assessment using deep learning: clinical implementation," *Radiology*, vol. 290, no. 1, pp. 52–58, 2018.

[171] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE T. Pattern Anal.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[172] G. Wu, M. Kim, Q. Wang, B.-C. Munsell, and D. Shen, "Scalable high-performance image registration framework by unsupervised deep feature representations learning," *IEEE T. Bio-med. Eng.*, vol. 63, no. 7, pp. 1505–1516, 2015.

[173] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241, Springer, 2015.

[174] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[175] T. P. Matthews, S. Singh, B. Mombourquette, J. Su, M. P. Shah, S. Pedemonte, A. Long, D. Maffit, J. Gurney, R. M. Hoil, *et al.*, "A multi-site study of a breast density deep learning model for full-field digital mammography and digital breast tomosynthesis exams," *arXiv preprint arXiv:2001.08383*, 2020.

[176] U. Nations, "Resolution adopted by the General Assembly on 28 june 2012." 66/281. International Day of Happiness, June 2012.

[177] I. Day of Happiness.

[178] U. Nations Foundation, "Pharrel Williams tells the world: A Happy Planet = Happy People." 2015.
[179] U. Nations, "International Day of Happiness.".
[180] U. Nations, "Sustainable Development Goals.".
[181] M.-E. Seligman and M. Csikszentmihalyi, "Positive psychology: An introduction," *Am. Psychol*, vol. 55, no. 1, pp. 5–14, 2000.
[182] D.-G. Myers, "The funds, friends, and faith of happy people.," *Am. Psychol.*, vol. 55, no. 1, p. 56, 2000.
[183] E. Diener, "The science of happiness and a proposal for a national index," *Am. Psychol.*, vol. 55, no. 1, pp. 34–43, 2000.
[184] M. Csikszentmihalyi, "If we are so rich, why aren't we happy?," *Am. Psychol.*, vol. 54, no. 10, p. 821, 1999.
[185] S. Lyubomirsky, L. Ki, and E. Diener, "The benefits of frequent positive affect: Does happiness lead to success?," 2005.
[186] R. Costanza, B. Fisher, S. Ali, C. Beer, L. Bond, R. Boumans, N. L. Danigelis, J. Dickinson, C. Elliott, J. Farley, *et al.*, "Quality of life: An approach integrating opportunities, human needs, and subjective well-being," *Ecol. Econ.*, vol. 61, no. 2, pp. 267–276, 2007.
[187] S. M. Schueller and M. E. Seligman, "Pursuit of pleasure, engagement, and meaning: Relationships to subjective and objective measures of well-being," *J. Posit. Psychol.*, vol. 5, no. 4, pp. 253–263, 2010.
[188] N. L. Sin and S. Lyubomirsky, "Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: A practice-friendly meta-analysis," *J. Clin. Psychol.*, vol. 65, no. 5, pp. 467–487, 2009.
[189] M. E. Seligman, "Positive health," *Applied Psychology*, vol. 57, no. s1, pp. 3–18, 2008.
[190] S. Joseph and C. A. Lewis, "The depression–happiness scale: Reliability and validity of a bipolar self-report scale," *J. Clin. Psychol.*, vol. 54, no. 4, pp. 537–544, 1998.
[191] S. Joseph, P.-A. Linley, J. Harwood, C.-A. Lewis, and P. McCollam, "Rapid assessment of well-being: The short depression-happiness scale (sdhs)," *Psychol. Psychother-T.*, vol. 77, no. 4, pp. 463–478, 2004.
[192] M. Rantanen, S. Mauno, U. Kinnunen, and J. Rantanen, "Do individual coping strategies help or harm in the work–family conflict situation? examining coping as a moderator between work–family conflict and well-being.," *Int. J. Stress Manage*, vol. 18, no. 1, p. 24, 2011.
[193] M. Ojala, "How do children cope with global climate change? coping strategies, engagement, and well-being," *J. Environ. Psychol.*, vol. 32, no. 3, pp. 225–233, 2012.
[194] L. Lu, J. Shih, Y. Lin, and L. Ju, "Personal and environmental correlates of happiness," *Pers. Indiv. Differ.*, vol. 23, no. 3, pp. 453–462, 1997.

[195] Y.-K. Chan and R. P.-L. Lee, "Network size, social support and happiness in later life: A comparative study of beijing and hong kong," *J. Happiness Stud.*, vol. 7, no. 1, pp. 87–112, 2006.

[196] F. Gülaçtı, "The effect of perceived social support on subjective well-being," *Proc. Soc. Behv.*, vol. 2, no. 2, pp. 3844–3849, 2010.

[197] K. L. Siedlecki, T. A. Salthouse, S. Oishi, and S. Jeswani, "The relationship between social support and subjective well-being across age," *Soc. Indic. Res.*, vol. 117, no. 2, pp. 561–576, 2014.

[198] A. Furnham and C.-R. Brewin, "Personality and happiness," *Pers. Indiv. Differ.*, vol. 11, no. 10, pp. 1093–1096, 1990.

[199] J. Brebner, J. Donaldson, N. Kirby, and L. Ward, "Relationships between happiness and personality," *Pers. Indiv. Differ.*, vol. 19, no. 2, pp. 251–258, 1995.

[200] N. Pishva, M. Ghalehban, A. Moradi, and L. Hoseini, "Personality and happiness," *Proc. Soc. Behv.*, vol. 30, pp. 429–432, 2011.

[201] L. Lu and J. Shih, "Personality and happiness: Is mental health a mediator?," *Pers. Indiv. Differ.*, vol. 22, no. 2, pp. 249–256, 1997.

[202] D. Campos, A. Cebolla, S. Quero, J. Bretón-López, C. Botella, J. Soler, J. García-Campayo, M. Demarzo, and R.-M. Baños, "Meditation and happiness: Mindfulness and self-compassion may mediate the meditation–happiness relationship," *Pers. Indiv. Differ.*, vol. 93, pp. 80–85, 2016.

[203] E.-M. Schmidt and Y.-E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 65–68, IEEE, 2011.

[204] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine.," in *Interspeech*, pp. 223–227, 2014.

[205] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Comput. Methods Programs Biomed.*, vol. 140, pp. 93–110, 2017.

[206] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint arXiv:1410.8586*, 2014.

[207] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[208] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," *arXiv preprint arXiv:1509.06041*, 2015.

[209] E. Cambria, B. Schuller, Y. Xia, and B. White, "New avenues in knowledge bases for natural language processing," *Knowledge-Based Systems*,

vol. 108, pp. 1 – 4, 2016. New Avenues in Knowledge Bases for Natural Language Processing.

[210] E. Cambria, E. Howard, Y. Xia, and T.-S. Chua, "Computational intelligence for big social data analysis [guest editorial]," *IEEE Comput. Intell. M.*, vol. 11, pp. 8–9, Aug 2016.

[211] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intell. Syst.*, vol. 31, pp. 102–107, Mar 2016.

[212] S. Poria, E. Cambria, and A.-F. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis.," in *EMNLP*, pp. 2539–2544, 2015.

[213] C.-N. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts.," in *COLING*, pp. 69–78, 2014.

[214] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng, "User-level psychological stress detection from social media using deep neural network," in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 507–516, ACM, 2014.

[215] X. Li, B. Hu, S. Sun, and H. Cai, "Eeg-based mild depressive detection using feature selection methods and classifiers," *Comput. Methods Programs Biomed.*, vol. 136, pp. 151–161, 2016.

[216] M. Argyle, "18 causes and correlates of happiness," *Well-being: The foundations of hedonic psychology*, vol. 353, 2003.

[217] J.-M. Bland and D.-G. Altman, "Statistics notes: Cronbach's alpha," *Brit, Med. J.*, vol. 314, no. 7080, p. 572, 1997.

[218] C.-S. Carver, "You want to measure coping but your protocol's too long: Consider the brief cope," *Int. J. Behav. Med*, vol. 4, no. 1, p. 92, 1997.

[219] L.-J. Francis, L.-B. Brown, and R. Philipchalk, "The development of an abbreviated form of the revised eysenck personality questionnaire (epqr-a): Its use among students in england, canada, the usa and australia," *Pers. Individ. Differ*, vol. 13, no. 4, pp. 443–449, 1992.

[220] D. Goldberg, "Manual of the ghq," *NFER, Windwor*, 1978.

[221] C.-D. Sherbourne and A.-L. Stewart, "The mos social support survey," *Soc. Sci. Med.*, vol. 32, no. 6, pp. 705–714, 1991.

[222] V.-E. Bonilla, "Confiabilidad, en el boletín informativo ineva en acción, vol. 2," 2006.

[223] I. Lucero and S. Meza, "Validación de instrumentos para medir conocimientos," *FACENA: Facultad de Ciencias Exactas y Naturales y Agrimesura de la UNNE*, 2002.

[224] A. Anastasi, "Psychological testing," 1968.

[225] I.-H. Bernstein and J.-C. Nunnally, "Psychometric theory, new york: Mcgraw-hill. oliva, ta, oliver, rl, & macmillan, ic (1992). a catastrophe model for developing service satisfaction strategies.," *J. Marketing*, vol. 56, pp. 83–95, 1994.

[226] R.-E. Bryant, "Symbolic boolean manipulation with ordered binary-decision diagrams," *ACM Comput. Surv.*, vol. 24, no. 3, pp. 293–318, 1992.

[227] S. Garavaglia and A. Sharma, "A smart guide to dummy variables: Four applications and a macro," in *Proceedings of the Northeast SAS Users Group Conference*, p. 43, 1998.

[228] J.-V. Tu, "Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes," *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, 1996.

[229] Z. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, no. 1-2, pp. 239–263, 2002.

[230] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma," *Neural Comput.*, vol. 4, no. 1, pp. 1–58, 1992.

[231] M.-I. Jordan *et al.*, "Why the logistic function? a tutorial discussion on probabilities and neural networks," 1995.

[232] B.-M. Wilamowski, "Neural network architectures and learning algorithms," *IEEE Ind. Electron. M.*, vol. 3, no. 4, 2009.

[233] S. Lawrence, C.-L. Giles, and A.-C. Tsoi, "Lessons in neural network training: Overfitting may be harder than expected," in *AAAI/IAAI*, AAAI/IAAI, pp. 540–545, 1997.

[234] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models.," in *The International Society of Music Information Retrieval*, pp. 621–626, 2009.

[235] P. Valdez and A. Mehrabian, "Effects of color on emotions.," *J. Exp. Psychol. Gen.*, vol. 123, no. 4, p. 394, 1994.

[236] P.-R. Goldin, C.-A.-C. Hutcherson, K.-N. Ochsner, G.-H. Glover, J.-D.-E. Gabrieli, and J.-J. Gross, "The neural bases of amusement and sadness: a comparison of block contrast and subject-specific emotion intensity regression approaches," *Neuroimage*, vol. 27, no. 1, pp. 26–36, 2005.

[237] C. Van Campen and J. Iedema, "Are persons with physical disabilities who participate in society healthier and happier? structural equation modelling of objective participation and subjective well-being," *Qual. Life Res.*, vol. 16, no. 4, p. 635, 2007.

[238] J. Zhang, D. Miao, Y. Sun, R. Xiao, L. Ren, W. Xiao, and J. Peng, "The impacts of attributional styles and dispositional optimism on subject well-being: A structural equation modelling analysis," *Soc. Indic. Res.*, vol. 119, no. 2, pp. 757–769, 2014.

[239] R.-M. Warner and D. Rasco, "Structural equation models for prediction of subjective well-being: Modeling negative affect as a separate outcome," *The Journal of Happiness & Well-Being*, vol. 2, no. 1, pp. 34–50, 2014.

[240] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*, vol. 1. MIT press Cambridge, 2016.

[241] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[242] V. Antonov, D. Tarkhov, and A. Vasilyev, "Unified approach to constructing the neural network models of real objects. part 1," *Math. Method. Appl. Sci.*, vol. 41, no. 18, pp. 9244–9251, 2018.

[243] J. Arifovic and R. Gencay, "Using genetic algorithms to select architecture of a feedforward artificial neural network," *Physica A: Statistical Mechanics and its Applications*, vol. 289, no. 3-4, pp. 574–594, 2001.

[244] B.-U. Islam, Z. Baharudin, M.-Q. Raza, and P. Nallagownden, "Optimization of neural network architecture using genetic algorithm for load forecasting," in *5th International Conference on Intelligent and Advanced Systems (ICIAS), 2014*, 5th International Conference on Intelligent and Advanced Systems (ICIAS), 2014, pp. 1–6, 2014.

[245] J. Koutník, J. Schmidhuber, and F. Gomez, "Evolving deep unsupervised convolutional networks for vision-based reinforcement learning," in *Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation*, Proceedings of the 2014 Annual Conference on Genetic and Evolutionary Computation, pp. 541–548, ACM, 2014.

[246] I. Loshchilov and F. Hutter, "Cma-es for hyperparameter optimization of deep neural networks," *arXiv preprint arXiv:1604.07269*, 2016.

[247] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," *arXiv preprint arXiv:1703.03864*, 2017.

[248] P. Vidnerová and R. Neruda, "Evolving keras architectures for sensor data analysis," in *Federated Conference on Computer Science and Information Systems (FedCSIS), 2017*, Federated Conference on Computer Science and Information Systems (FedCSIS), 2017, pp. 109–112, IEEE, 2017.

[249] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys*, vol. 74, no. 1, p. 47, 2002.

[250] M. Newman, A.-L. Barabási, and D.-J. Watts, *The structure and dynamics of networks*, vol. 19. Princeton University Press, 2011.

[251] R. Albert, H. Jeong, and A.-L. Barabási, "Internet: Diameter of the world-wide web," *Nature*, vol. 401, no. 6749, p. 130, 1999.

[252] S. Redner, "How popular is your paper? an empirical study of the citation distribution," *Eur. Phys. J. B*, vol. 4, no. 2, pp. 131–134, 1998.

[253] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proc. Natl. Acad. Sci. U.S.A*, vol. 98, no. 8, pp. 4569–4574, 2001.

[254] A.-L. Barabási, "Network medicine - from obesity to the diseasom," 2007.

[255] F. Jian and S. Dandan, "Complex network theory and its application research on p2p networks," *Applied Mathematics and Nonlinear Sciences*, vol. 1, no. 1, pp. 45–52, 2016.

[256] A.-L. Barabási and M. Pásfai, *Network science*. Cambridge: Cambridge University Press, 2016.

[257] S. Fortunato and C. Castellano, "Community structure in graphs," in *Computational Complexity*, Computational Complexity, pp. 490–512, Springer, New York, NY., 2012.

[258] B.-W. Kernighan and S. Lin, "An efficient heuristic procedure for partitioning graphs," *Bell Syst. Tech. J.*, vol. 49, no. 2, pp. 291–307, 1970.

[259] J. Scott, *Social network analysis*. Sage Publications, 2017.

[260] L.-A.-N. Amaral, A. Scala, M. Barthelemy, and H.-E. Stanley, "Classes of small-world networks," *Proc. Natl. Acad. Sci. U.S.A*, vol. 97, no. 21, pp. 11149–11152, 2000.

[261] M. Marchiori and V. Latora, "Harmony in the small-world," *Physica A: Statistical Mechanics and its Applications*, vol. 285, no. 3-4, pp. 539–546, 2000.

[262] W. Luo, N. Lu, L. Ni, W. Zhu, and W. Ding, "Local community detection by the nearest nodes with greater centrality," *Information Sciences*, 2020.

[263] P. Yanardag and S.-V.-N. Vishwanathan, "Deep graph kernels," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1365–1374, 2015.

[264] T.-N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[265] J. Li, H. Zhang, Z. Han, Y. Rong, H. Cheng, and J. Huang, "Adversarial attack on community detection by hiding individuals," *arXiv preprint arXiv:2001.07933*, 2020.

[266] M. Khodayar and J. Wang, "Spatio-temporal graph deep neural network for short-term wind speed forecasting," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 2, pp. 670–681, 2018.

[267] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[268] P.-E. Spector, *Summated rating scale construction: An introduction*, vol. 82. Sage Publications, 1992.

[269] J.-T. Cacioppo and G.-G. Berntson, "Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates.," *Psychol. Bull*, vol. 115, no. 3, p. 401, 1994.

[270] A. Clauset, M.-E.-J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, p. 066111, 2004.

[271] D.-S. Wilks, "Cluster analysis," in *International geophysics*, vol. 100, pp. 603–616, Elsevier, 2011.

[272] A. Arenas, J. Duch, A. Fernández, and S. Gómez, "Size reduction of complex networks preserving modularity," *New J. Phys*, vol. 9, no. 6, p. 176, 2007.

[273] D.-P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[274] V.-A. Traag, "Faster unfolding of communities: Speeding up the louvain algorithm," *Physical Review E*, vol. 92, no. 3, p. 032801, 2015.

[275] A. Hagberg, P. Swart, and D. S Chult, "Exploring network structure, dynamics, and function using networkx," tech. rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[276] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.