Mathematical modelling and learning of biomedical signals for safety pharmacology

Fabien Raphel

Mathematical modelling and learning of biomedical signals for safety pharmacology.

THÈSE DE DOCTORAT

Présentée par

Fabien Raphel

pour obtenir le grade de

DOCTEUR DE
Sorbonne Université

Spécialité : Mathématiques Appliquées

Soutenue publiquement le JJ MMMMMM AAAA devant le jury composé de :

Directeur de Thèse	Sylvain BERNASCONI	Directeur de Notocord
########	Albert COHEN	Professeur de l'UPMC Paris 6
Directeur de Thèse	Jean-Frédéric GERBEAU	DGDS de l'Inria
Co-directeur de Thèse	Damiano LOMBARDI	Chargé de recherche de l'Inria Paris
Rapporteur	Gary MIRAMS	Professeur de l'Université de Nottingham
########	Molly MALECKAR	Professeure à Simula Oslo
Rapporteur	Rodolphe TURPAULT	Professeur de Bordeaux INP
########	Karen VEROY	Professeure de l'Université de Eindhoven

Après avis favorables des rapporteurs: Gary MIRAMS et Rodolphe Turpault

Thèse préparée au sein de l'équipe-projet Commedia
Centre de Recherche Inria de Paris
2 rue Simone Iff
75589 Paris Cedex 12
et **Laboratoire Jacques-Louis Lions**
de **Sorbonne Université**

Résumé:

En tant que branche de la pharmacologie, la pharmacologie de sécurité cardiaque vise à étudier les effets secondaires des composés sur le système cardiaque, à des doses thérapeutiques. Ces études, réalisées par le biais d'expériences *in silico*, *in vitro* et *in vivo*, permettent de sélectionner/rejeter un composé à chaque étape du processus de développement du médicament. Un vaste sous-domaine de la pharmacologie de sécurité cardiaque est consacré à l'étude de l'activité électrique des cellules cardiaques à partir d'expériences *in silico* et *in vitro*. Cette activité électrique est la conséquence d'échanges de structures polarisées (principalement des ions) entre le milieu extracellulaire et intracellulaire. Une modification des échanges ioniques induit des changements dans l'activité électrique de la cellule cardiaque qui peuvent être pathologiques (par ex. en générant des arythmies). Une bonne connaissance de ces signaux électriques est donc essentielle pour prévenir les risques d'évènements létaux.

Les techniques de patch-clamp sont les méthodes les plus courantes pour enregistrer l'activité électrique d'une cellule cardiaque. Bien que ces signaux électriques soient bien connus, ils sont lents et fastidieux à réaliser, et donc, coûteux. Une alternative récente consiste à considérer les dispositifs de réseaux de microélectrodes (MEA). Développés à l'origine pour l'étude des neurones, leur extension aux cellules cardiaques permet un criblage à haut débit qui n'était pas possible avec les techniques de patch-clamp. Un MEA est une plaque avec des puits dans lesquels des cellules cardiaques (formant un tissu) recouvrent des électrodes. Par conséquent, l'extension de ces dispositifs aux cellules cardiaques permet d'enregistrer l'activité électrique des cellules au niveau du tissu (avant et après l'ajout d'un composé dans les puits). Comme il s'agit d'un nouveau signal, de nombreuses études doivent être menées pour comprendre comment les échanges ioniques induisent cette activité électrique enregistrée, et, enfin, pour procéder à la sélection/rejet d'un composé. Bien que ces signaux soient encore mal connus, des études récentes ont montré des résultats prometteurs dans la prise en compte des MEA dans la pharmacologie de sécurité cardiaque. L'automatisation de la sélection/rejet d'un composé est encore difficile et loin des applications industrielles, ce qui est l'objectif final de ce manuscrit.

Mathématiquement, le processus de sélection/rejet peut être considéré comme un problème de classification binaire. Comme dans toute classification supervisée (et dans les tâches d'apprentissage automatique, plus généralement), une entrée doit être définie. Dans notre cas, les séries temporelles des activités électriques cardiaques sont éventuellement longues (minutes ou heures) avec un taux d'échantillonnage élevé ($\sim kHz$) conduisant à une entrée appartenant à un espace de grande dimension (centaines, milliers ou même plus). De plus, le nombre de données disponibles est encore faible (au plus quelques centaines). Ce régime critique nommé haute dimension/faible taille d'échantillon rend le contexte difficile. Le but de ce manuscrit est de fournir une stratégie systématique pour sélectionner/rejeter des composés d'une manière automatisée, sous les contraintes suivantes:

- Traiter le régime de haute dimension/faible taille d'échantillon.

- Aucune hypothèse sur la distribution des données.
- Exploiter les modèles *in silico* pour améliorer les performances de classification.
- Pas ou peu de paramètres à régler.

La première partie du manuscrit est consacrée au contexte, suivie de la description des techniques de patch-clamp et de MEA. Enfin, une description des modèles de potentiel d'action et de potentiel de champ pour réaliser des expériences *in silico* est donnée.

Dans une seconde partie, deux aspects méthodologiques sont développés en respectant au mieux les contraintes définies par le contexte industriel. Le premier décrit une stratégie de réduction de l'espace d'entrée basée sur une fonction score liée au taux de succès de la classification. Des comparaisons avec des méthodes classiques de réduction de dimension telles que PCA et PLS (avec leurs paramètres par défaut) sont effectuées, montrant que la méthode proposée conduit à de meilleurs résultats. La deuxième méthode consiste en la construction d'un ensemble d'entraînement augmenté basé sur un réservoir de simulations, en considérant la distance de Hausdorff entre les ensembles et la maximisation de la même fonction score que pour la première méthode. La stratégie proposée permet de rejeter automatiquement les données biaisées et/ou mal étiquetées pour construire l'ensemble d'entraînement augmenté. Une expérience numérique est réalisée sur des potentiels d'action *in silico* et une comparaison avec SVM et KNN (avec leurs paramètres par défaut) est effectuée, montrant que la méthode proposée conduit globalement à des taux de succès plus élevés.

Dans la troisième partie, deux applications sur des données de patch-clamp sont réalisées. Dans une première étude, il s'agit d'un problème de régression pour estimer l'activité des canaux ioniques à partir de signaux de potentiel d'action *in silico*. Le couplage de la méthode de réduction de dimension orientée avec un filtre de Kalman améliore la qualité de l'estimation de l'activité du canal ionique en terme de temps de calcul et précision. Une deuxième étude est consacrée à la classification (de la forme impact sur le signal électrique Oui/Non) de composés à partir de signaux de patch-clamp automatisés. La méthode proposée pour la réduction de dimension à conduit à une meilleure classification des composés (en particulier à des concentrations intermédiaires) qu'avec la stratégie de classification proposée par la société avec laquelle nous avons collaborée.

Enfin, la quatrième partie du manuscrit est consacrée aux signaux de MEA. Dans une première étude, une méthode de réduction de dimension orientée est appliquée à des expériences *in vitro* montrant des résultats qualitativement bons de classification des canaux ioniques. La deuxième étude est dédiée au couplage des deux méthodes proposées. En particulier, l'application aux données de la première conduit à l'amélioration du taux de succès en utilisant la stratégie de construction d'un ensemble d'entraînement augmenté. Cette partie se termine par une application sur un plus grand ensemble de données est effectuée pour considérer les deux mèthodes proposées dans un contexte industriel.

Abstract: As a branch of pharmacology, cardiac safety pharmacology aims at investigating compound side effects on the cardiac system at therapeutic doses. These investigations, made through *in silico*, *in vitro* and *in vivo* experiments, allow to select/reject a compound at each step of the drug development process. A large subdomain of cardiac safety pharmacology is devoted to the study of the electrical activity of cardiac cells based on *in silico* and *in vitro* assays. This electrical activity is the consequence of polarised structure exchanges (mainly ions) between the extracellular and intracellular medium. A modification of the ionic exchanges induces changes in the electrical activity of the cardiac cell which can be pathological (e.g. by generating arrhythmia). Strong knowledges of these electrical signals are therefore essential to prevent risk of lethal events.

Patch-clamp techniques are the most common methods to record the electrical activity of a cardiac cell. Although these electrical signals are well known, they are slow and tedious to perform, and therefore, expansive. A recent alternative is to consider microelectrode array (MEA) devices. Originally developped for neurons studies, its extension to cardiac cells allows a high throughput screening which was not possible with patch-clamp techniques. It consists of a plate with wells in which cardiac cells (forming a tissue) cover some electrodes. Therefore, the extension of these devices to cardiac cells allow to record the electrical activity of the cells at a tissue level (before and after compound addition into the wells). As a new signal, many studies have to be done to understand how ionic exchanges induce this recorded electrical activity, and, finally, to proceed the selection/rejection of a compound. Despite these signals are still not well known, recent studies have shown promising results in the consideration of MEA into cardiac safety pharmacology. The automation of the compound selection/rejection is still challenging and far from industrial applications, which is the final goal of this manuscript.

Mathematically, the selection/rejection process can be seen as a binary classification problem. As in any supervised classification (and machine learning tasks, more generally), an input has to be defined. In our case, time series of the cardiac electrical activities are possibly long (minutes or hours) with a high sampling rate ($\sim kHz$) leading to an input living in a high-dimensional space (hundreds, thousands or even more). Moreover the number of available data is still low (at most hundreds). This critical regime named high dimension/low sample size make the context challenging.

The aim of this manuscript is to provide a systematic strategy to select/reject compounds in an automated way, under the following constraints:

- Deal with high dimension/low sample size regime.
- No assumptions on the data distributions.
- Exploit *in silico* models to improve the classification performances.
- No or few parameters to tune.

The first part of the manuscript is devoted to the context, followed by the description of the patch-clamp and MEA technologies. This part ends by the description of action potential and field potential models to perform *in silico* experiments.

In a second part, two methodological aspects are developped, trying to comply, at best, with the constraints of the industrial application. The first one describes a double greedy goal-oriented strategy to reduce the input space based on a score function related to the classification success rate. Comparisons with classical dimension reduction methods such as PCA and PLS (with default parameters) are performed, showing that the proposed method led to better results. The second method consists in the construction of an augmented training set based on a reservoir of simulations, by considering the Hausdorff distance between sets and the maximisation of same score function as in the first method. The proposed strategy makes it posssible to automatically reject biased and/or wrongly labelled data to construct the augmented training set. A numerical experiments is performed on *in silico* action potentials and comparison with SVM and KNN (with default parameters) are done, showing that the proposed method globally led to higher classification success rates.

In the third part, two applications to patch-clamp data are performed. In a first study, it consists in a regression problem to estimate the ion channel activity from *in silico* action potential signals. The coupling of the proposed goal-oriented double greedy dimension reduction method with an unscented Kalman filter improves the ion channel activity estimation in terms of computational cost and accuracy. A second study is devoted to the Hit/No hit classification of compounds based on automated patch-clamp signals. The goal-oriented dimension reduction methods led to a better classification of the compounds (particularly at intermediate concentrations) than with the classification strategy proposed by the company with whom we collaborated.

Finally, the part number four of the manuscript is devoted to MEA signals. In a first study, a goal-oriented double greedy dimension method is applied to *in vitro* experiments showing qualitatively good ion channel classification results. The second study highlights the improvements of the classification success rate using the augmented training set construction strategy. This part ends by an application on a larger dataset is performed to consider the two proposed methods into an industrial context.

Contents

List of Figures

List of Tables

Part I

Introduction

Chapter 1

Preamble

Contents

1.1 Context: Safety pharmacology

Drug development is the process starting from the identification of a molecule with a potential to become a therapeutic agent (drug) and placing it on the market (accessibility to the patient). This long process (around 10 years) consists in pruning a large number of candidates to keep the desired ones. Drug development can be divided into different parts, as presented in Figure 1.1.

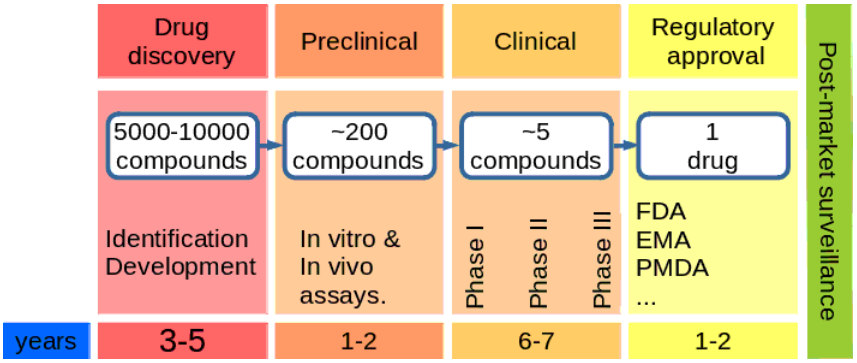

Figure 1.1: Phases of the drug development process.

The number of approved drugs by the regulation authority (European Medicines Agency, Food and Drug Administration, Pharmaceuticals Medical Devices Agency, ...) is then very low with respect to the number of candidates at preclinical and clinical stages [HTC+14, WSL19] and can even drop to less than 10^{-2}% from drug discovery to regulatory approval [YKN21]. Improving the compound selection is then crucial to reduce the delays between the first stage and the regulatory approval.

At preclinical steps, the selection process consists in the realisation of *in vitro* and *in vivo* assays. Physiological responses under compound addition guide the pharmacologist to select or reject the compound to prevent lethal events at clinical steps. As a branch of pharmacology, **safety pharmacology** aims at selecting/rejecting compounds by predicting the risk of rare lethal deaths, based on *in silico*, *in vitro* and *in vivo* assays at therapeutic concentrations [PAC08, RWP+02].

In this manuscript we will focus on the cardiac part of the safety pharmacology, the cardiotoxicity, and more precisely on *in vitro* experiments. Cardiac safety pharmacology is devoted to the evaluation of the effects (undesirable or not) of a given candidate at a given concentration on the cardiovascular system, before clinical trials. For *in vitro* assays, these experiments are performed on cultured cardiac cells which can be isolated or forming a functional syncytium (i.e. a cardiac tissue). Along with these experimental studies, *in silico* models emerged to better understand the cardiac cell functionality and improve disease forecasting.

Guidelines such as the International Conference of Harmonisation (ICH) [Bra05, CFG+16] allow a standardisation of the drug development process. It consists in preventing fatal effects (such as sudden death caused by arrhythmia[1]) during clinical steps by systematically removing compound inducing targeted effects specified by the guidelines.

As an example, hERG[2] inhibitors are known to prolong the QT interval[3] which is related

[1] Cardiac rhythm abnormality.

[2] hERG is the acronym of the human Ether-a-go-go-Related Gene corresponding to a potassium ion channel. See Section 2.2 for more details on the cardiac cell.

[3] QT is the duration between the beginning of the Q peak and the end of the T wave in an electrocar-

to arrhythmia (and Torsades de Pointes [RPM+05] more precisely). For these reasons, hERG blockers are removed at early stages. However, some of these removed entities might potentially become drugs [WBD+18]. Therefore this causality between QT prolongation and Torsades de Pointes risk seems not to be established and other quantities (i.e. biomarkers) might be considered to detect possible arrhythmia [Hon18]. This example highlights some limitations of the actual guidelines.

To go deeper into this investigation and avoid removing by excessive caution potential interesting compounds, several international consortiums such as Comprehensive in vitro Proarrhythmia Assay (CiPA) emerged. This new paradigm is based on a standardisation of the protocols introducing high throughput screening devices (Automated patch-clamp and MEA[4]), and aims at going deeper than the ICH guidelines by:

- Defining more precise biomarkers to better quantify arrhythmia risk.
- Identifying ion channel blockade.
- Considering *in silico* models.

These investigations are essential to improve the pruning steps in the life cycle of drug development (see Figure 1.1). For a given compound it corresponds to a series of tests for which each step can be seen as a "yes/no" answer to a question summarised as "Can we go to the next step with this compound?". Each step is associated with a specific protocol leading to an information (e.g. a physiological signal) helping to answer it. Physiologically speaking, this question can be reformulated as "Does this compound at a given concentration can induce arrhythmia?" or "Does this compound at a given concentration blocks the potassium channel?". Other way stated, each compound at a given concentration has to be categorised based on the induced physiological response.

1.2 Problematic and mathematical aspects

In this manuscript, we described and investigated mathematical methods, which in the form of a tool, help the pharmacologist to decide whether a compound can pass to the next step of the drug development.

1.2.1 Classification problem

The problems we are dealing with are of the form: *"This compound at this concentration may induce arrhythmia"* or *"This compound at this concentration is a high blocker of this ion channel"* depending on the raised question and given an observation g_i. These kinds of yes/no answers fall in the class of supervised classification problems, a topic of applied mathematics and machine learning. Definition 1 describes a classification problem.

Definition 1
Let $T_g^{n_s} = \{g_i \in \Omega \subseteq \mathbb{R}^{n_g}\}$ be a set of n_s observed elements ($i = 1, \dots, n_s$), paired with $T_y^{n_s} = \{y_i \in \Omega_y \subseteq \mathbb{F} = \mathbb{N} \text{ or } \mathbb{R}\}$ output elements. Given a new observation $\tilde{g} \in \Omega$, the goal of the classification problem is to determine its unknown output $\tilde{y} \in \Omega_y$ based on the knowledge of the labelled observations $(T_g^{n_s}, T_y^{n_s})$.

diogram (electrical signals of the heart recorded at specific regions of the body) corresponding to the delay between the cardiac cell depolarisation and the repolarisation.

[4]MEA: Microelectrodes array.

Other way stated, each observation g_i is paired with its corresponding output y_i. In a yes/no classification problem, we then have $\Omega_y = \{0,1\}$ (with "yes"= 0 and "no"= 1 or the reverse). The extension to more classes is trivial. In the case where the output belongs to $\mathbb{R}$ ($\mathbb{F} = \mathbb{R}$) the problem is a regression problem.

When the output is unknown, i.e. $T_y^{n_s}$ does not exist, the problem falls in the unsupervised classification (or clustering) branch of machine learning.

Then, in a supervised classification problem context, the experimenter has to give some data paired with a labelled set corresponding to the question to be answered. This labelled database is called Training set in the machine learning community. The first challenge is then to construct a classifier which has learnt to discriminate the different classes through the given Training set. Then, mathematically, a classifier is seen as a function with an observation as an input and the predicted class as an output, which is formalised as:

Definition 2
Let $g \in \Omega \subseteq \mathbb{R}^{n_g}$ be an observation, paired with a label $y \in \Omega_y \subseteq \mathbb{N}$. A classifier is a function $\mathcal{C}$ such that the following holds:

$$\begin{array}{l} \mathcal{C} : \Omega \to \Omega_y \\ g \to y \end{array},$$

where $Card(\Omega_y)$ corresponds to the number of classes.

We now have to define an "observation". In the context of this manuscript, we study signals varying in time (i.e. time series) as a signature of the cardiac cells activities (see Section 2 for more details). Let $\mathcal{G}$ be a set of these signals. For each $\mathcal{G}^{(i)}$, $i = 1, \ldots, n_s$ a set of $n_g \in \mathbb{N}^*$ quantities is extracted (e.g. amplitudes, durations, wavelet coefficients, ...). An observation is then defined as follows:

Definition 3
Let $\mathcal{G}^{(i)} \in \mathbb{R}^{n_T}$ be a signal (time series of n_T elements in our case). Let f_D be a function defined as follows:

$$\begin{array}{l} f_D : \mathbb{R}^{n_T} \to \Omega \subseteq \mathbb{R}^{n_g} \\ \mathcal{G}^{(i)} \to g \end{array}.$$

We name by "observation" the output g of the function f_D.

The set G of the whole observations (i.e. for each element of $\mathcal{G}$) is named the dictionary. It follows that the observation corresponding to the i^{th} signals is $g = G^{(i)}$.

1.2.2 Challenges

As mentioned above, the signals studied in this manuscript are time series corresponding to recordings of the electrical cardiac cell activity. These time series have a minimal duration which is about $250ms$ (it can change according to the cardiac cell type) [YBS12] and can reach up to several minutes or even hours.

Indeed, no drug has an immediate effect on ion currents and therefore, a delay is necessary to observe a possible stabilised effect which leads to a recording of several cardiac beats in a row [BSV+17, GLG+15].

With a sampling rate between $20kHZ$ to $40kHZ$ [JKT+03, XWL14], it results in at least 5.10^3 recorded amplitudes for one cardiac beat (meaning that $\mathcal{G}^{(i)} \in \mathbb{R}^{n_T}$, with $n_T = 5.10^3$, for $i = 1, \ldots, n_s$). From $\mathcal{G}$ we now have to construct the function f_D to build G the dictionary matrix. Due to the large amount of recorded information, a large number of quantities has to be extracted in order to limit the loss of information carried by the signal. It follows that an observation g belongs to a high-dimensional space $\mathbb{R}^{n_g}$ (tens, hundreds, thousands or even more).

To deal with a supervised classification problem, we need observations. However, in many industrial applications (and in the present context of *in vitro* assays), the experiments which have to be performed to get these observations are time consuming and costly. In these circumstances, a few number of samples are then available to resolve the classification problem.

The input of the classification is therefore in a high-dimensional space ($\mathbb{R}^{n_g}$) whereas the number of observations n_s is quite low with respect to n_g.

This particular context named high dimension/low sample size is related to the "curse of dimensionality" introduced by Bellman [Bel15]. This critical regime tends to introduce overfitting (illustration and explanation are given in the left panel of Figure 1.2).

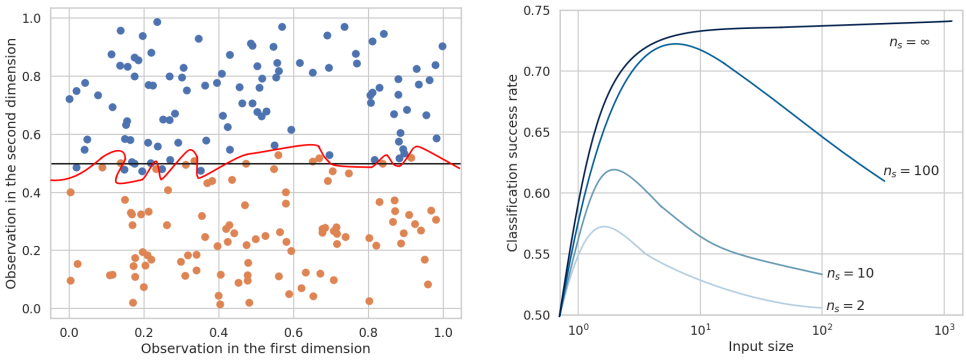

Figure 1.2: Left panel: Example of overfitting in $\mathbb{R}^2$. Blue and orange dots are labelled data (each colour corresponds to one class) providing from the Training set. The true delimitation is the horizontal black line. The class delimitation learnt by the classifier is the red line, meaning that observation above the red line will be classified "blue" whereas observation below the red line, observation will be classified "orange". Here, the classifier is too precise and has learnt the noise on the labelled data. It will result in bad classifications for observed data around the true delimitation line (black line). Right panel: Illustration of the Hughes phenomenon for different sample sizes n_s.

Even worse, for a given sample size n_s, the classifier performances will decrease as the dictionary entry size n_g will increase. This behaviour is known as Hughes phenomenon [Tad98] and is reported into the right panel of Figure 1.2.

Considering the above aspects, the main objective is to provide a set of tools which is able to:

- Deal with high dimension/low sample size regimes.
- Consider low/none assumptions on the observable space (i.e. no *a priori* on the probability distributions).
- Exploit *in silico* models to improve the classification performances.
- Have only a few parameters/hyperparameters to be tuned.

Being challenging for pharmacologists in terms of precision and timesaver, the tool must in addition be able to fit into an industrial framework.

1.3 Contributions

The work reported in this manuscript originates from the common laboratory (LabCom) CardioXcomp that has brought together NOTOCORD® part of Instem company and Inria (Carmen and Reo teams) started in October 2013 and ended in September 2016. This LabCom was devoted to mathematical modelling for pharmaceutical research and aimed at improving measurement devices (in particular microelectrode arrays; see Section 2.3.2) for cardiac cells more precisely. CardioXcomp project led to several contributions [BRGZ15, CGB+16, RBZ+17, ABC+18, TRLG18]. These works contributed to a better understanding of MEA modelling based on (human induced pluripotent stem cells derived to) cardiac cells (see Section 2.2 for more details) and compound effects on recorded signals. Several recent works followed these contributions to a better comprehension of MEA signals and its interests in cardiac safety assessments [JCW+20, Küg20].

In the spirit of these studies, the works presented in this manuscript are interdisciplinary between mathematics and pharmacology, and was carried out jointly at Inria and NOTOCORD®, part of Instem company. Its final objective is the introduction of the proposed methods in NOTOCORD's software. It would not have been possible without the vision of NOTOCORD's founder, Philippe Zitoun, and his collaborators, and without the numerous collaborations they have made possible with various companies and pharmacologists.

Thanks to Celine Hechard, we realised a first collaboration with Tessa de Korte and Stefan Braam, members of Ncardia[5]. Electrogram signals provided by microelectrode arrays (MEA) are studied without and under compound addition. Several quantities (biomarkers) were extracted from these signals. A method was proposed to construct a classifier based on a goal-oriented dimension reduction. A first application to Torsades de Pointes risk was applied on *in silico* experiments and led to a classification success rate close to 0.94. A second application based on *in silico* and *in vitro* MEA experiments was performed to assess ion channel blockade. It results in a classification success rate close to 0.89 for the potassium channel blockade. Classification results obtained for each tested compound is shown as an example in Figure 1.3.

This collaboration led to a published paper in PLOS Computational Biology [RDKL+20] and is reported into Section 13.

Despite this first study ended by encouraging results, several questions arose:

1. To perform the proposed strategy, a classification method has to be chosen (e.g. linear discrimant analysis or support vector machines). How to construct a method, which does not depend on the choice of the classifier?

   To answer this question, we investigated a new dimension reduction strategy:

   - Because of the high dimensional - low sample size context presented above, a dimension reduction strategy is needed to construct the input of the classification. Instead of using classical dimension reduction methods such as Feature Selection or Principal Componant Analysis, a **goal-oriented** strategy is proposed. This goal-oriented strategy aims at reducing the dimension while maximising the classification success rate (by considering the most relevant entries of the dictionary G), without depending

[5] Leiden, Netherlands. ncardia.com

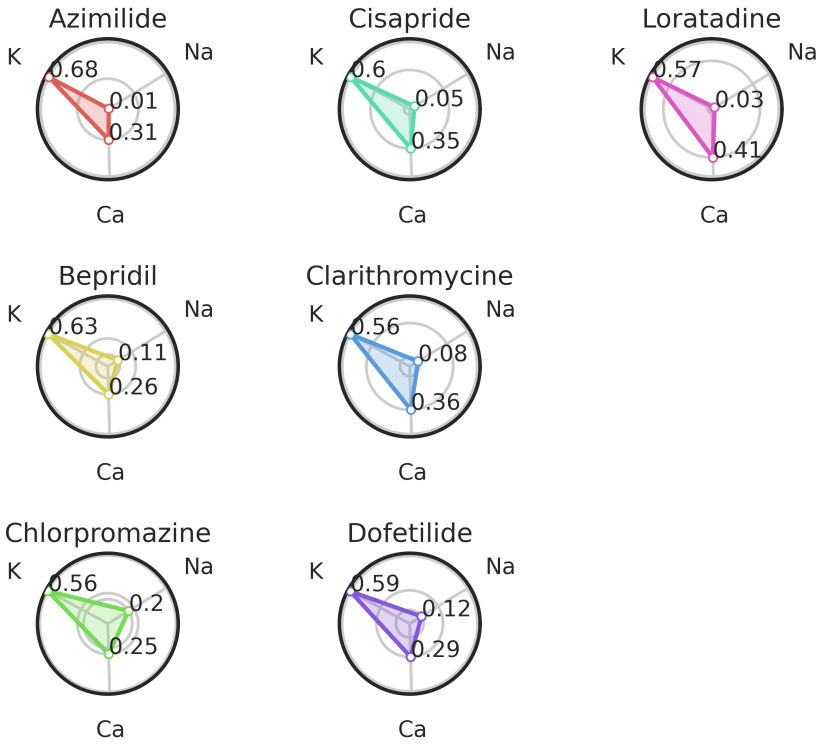

Figure 1.3: Compounds classification results obtained in Section 13.3.2.3. This ternary classification task aims at detecting which ion channel (potassium: K, sodium: Na, or calcium: Ca) is blocked by a compound. Values returned by the classifier (black values in the polar plots) are the probabilities to block the corresponding channel blocker.

on the choice of a classifier. Thus, a score function related to the classification success rate was introduced. This score function was compared with distances and f-divergences usually considered in classification problems. Instead of maximising the score function over the whole dictionary matrix, which can be challenging, a Double Greedy Dimension Reduction (DGDR) strategy is proposed to approximate the solution of the problem. Numerical studies were performed (with $n_g = 10^5$) to compare the proposed method with existing dimension reduction methods such as Neighbourhood Component Analysis and Partial Least Squares. Comparisons with existing dimension reduction methods were also made on experimental datasets. This study led to an submitted paper [LR19] corresponding to Section 5 in the manuscript.

The proposed Double Greedy Dimension Reduction (DGDR) method was then applied in two different occasions:

- A joint work with Esther Pueyo, David Adolfo Sampedro-Puente, Jesus Fernandez-Bes and Pablo Laguna from Zaragoza University aimed at improving ionic channel

activity estimation based on Action Potential signals[6], combining Unscented Kalman Filter method and the above algorithm extended to regression problems. Numerical experiments showed that the goal-oriented dimension reduction method extended to regression problems and coupled with UKF method improve the estimation of the ionic channel activities. This work resulted into a published paper into IEEE Journal of Biomedical and Health Informatics [SPRFB+20] corresponding to Section 9.

- A third collaboration was performed with Kadla Røskva Rosholm, Lasse Homann and Anders Lindqvist, members of Sophion Bioscience[7] to study compound effects on Nav1.7 channels based on automated patch-clamp signals. The aim of the project was to automate detection of Nav1.7 channel modulators, thereby increasing speed and accuracy of large compound screens. This detection consists in studying whether the recorded electrical signal (patch-clamp signal) is affected by the compound at a given concentration. This hit/no hit binary classification led to a classification success rate close to 0.95. The application of the DGDR method showed improvements of the classification performances if compared to the classification method currently used at Sophion. In particular, for the second-highest concentrations, the accuracy obtained with the DGDR method is 0.73 against 0.52 with the current evaluation used at Sophion.

  This work is reported into Section 10 of the manuscript. The collaboration is still in progress in view of a possible publication.

The relevance of the labelled data, whether it be the training or the validation set is essential for the dimension reduction process (whatever it is). How to deal with wrongly labelled data or data badly recorded (e.g. due to noise)? In addition, in several applications, simulations can be considered to enrich the database. This point can particularly be crucial in cases of costly experiments. However, how to embed these simulated experiments to improve classification performances? How to deal with simulations presenting a bias? Is the bias penalising the answer of the question raised by the experimenter?

To address these questions, a strategy was proposed to automatically construct an augmented training set:

- A methodological work was performed in order to construct an augmented training set. It consists in selecting the most relevant samples (with respect to a score function) to construct the training set. An algorithm called ASE-HD based on the Hausdorff distance between sets is proposed. Some theoretical aspects such as convergence were highlighted. Numerical examples were performed on three study cases. An application on *in silico* action potentials was realised in dimension $n_g = 24$ considering different levels of bias. This work led to a published paper into ESAIM:M2AN [LR21] and corresponds to Section 6.

2. While the DGDR method deals with the selection of the most relevant (with respect to the raised question) entries, the second methodological work deals with the selection of the most relevant samples (with respect to the same question). Roughly speaking, the first method works on the rows of the dictionary matrix G, while the second one works on the columns. How these two methodological works can be combined together to provide a framework? Two studies were considered to manage this coupling:

[6] Action Potentials are described in Section 2.2.

[7] Ballerup, Denmark. sophion.com

- The consideration of the two methodological parts was performed to improve results obtained in the works realised with Ncardia (first paper). Starting from a classification success rate close to 0.89 for the potassium channel blockade on a blind test set, the combination of the two proposed methods reached the classification success rate to 0.98 on this same set. These results are reported in Section 14.
- The last collaboration was perfomed with Udo Kraushaar from NMI[8]. This work aims at being the closest as possible to a final industrial application before integration into a software. This is still an ongoing work.

The methodological aspects described above are general and can be applied to other classifications problems despite the fact that it is applied to cardiac safety pharmacology. The above contributions were presented in the historical order of the works and raised questions. The structure of the manuscript does not follow this timeline but is partitioned following the parts described in the next section.

1.4 Organisation of the manuscript

The manuscript is partitioned into 4 main parts and resumes the above collaborations and contributions:

- The first part is devoted to the introduction of the problematic and the description of the technical aspects needed for the reading of the manuscript:
  - A preamble to introduce the problematic and the challenges.
  - A cardiac safety pharmacology section devoted to the cardiac cell functionality and measurement tools to record electrical activities.
  - A mathematical modelling and simulations part to perform *in silico* experiments.
- The second part (see Part II) describes the two methods to answer the problematic described above:
  - A first section describes the goal-oriented dimension reduction method (DGDR algorithm).
  - A second section deals with the construction of an augmented training set based on synthetic data, generated by numerical simulations (ASE-HD algorithm).
- In the third part, two applications on cardiac action potentials are performed (see Part III):
  - Regression problem to detect ion channel activity under β-adrenergic stimulation.
  - Hit/no hit classification based on automated patch-clamps signals.
- Finally, the last part is devoted to the application to microelectrode arrays (see Part IV):
  - Compound classification using the first version of the dimension reduction method.
  - Coupling of the two proposed methodological works and validation on experimental data.

An additional part is dedicated to the Conclusion (see Part V).

[8]Reutlinger, Germany. nmi.de

Chapter 2

Cardiac Safety Pharmacology

Cardiac safety pharmacology is a discipline appearing at preclinical phases of a drug development. Its main objective is to investigate whether a molecule could be a good candidate to become a drug, by detecting potential undesirable pharmacodynamic effects (such as arrhythmia) at therapeutic range.

In vitro assays allow a preliminary pruning of the candidates before *in vivo* assays. In cardiac safety pharmacology, these *in vitro* assays are performed on cultured cardiac cells. Its electrical properties being closely related to its contractility, several techniques and devices were developed to measure it such as patch-clamp or microelectrode arrays.

Contents

2.1 Introduction

Safety pharmacology is a discipline which aims at evaluating risk/benefit of molecules in a drug discovery context [PAC08, PHdK+18]. Each molecule is studied on the heart to check whether it has an undesirable effect on it. This study falls in the cardiac safety pharmacology branch, whose assays mainly deal with hemodynamics and electrophysiological experiments.

To study these molecule effects, as in any physical problems, two tools are needed: the object we want to observe and how we observe it (the measurement tool). In this manuscript, we will only focus on the early stages of cardiac safety pharmacology: *in vitro* assays. At this stage, no assays on animals (*in vivo*) are performed yet, meaning that experiments are achieved on cultured cardiac cells (described in Section 2.2). The main characterisation of a cardiac cell is its contractility induced by its electrical activity. This is why different techniques and devices were developed to measure this electrical activity (see Section 2.3).

This part is divided into two sections. In Section 2.2 a coarse description of the cardiac cell is given with more details on the origin of the electrical activity. In Section 2.3 main techniques and tools to measure this electrical activity are presented.

2.2 Cardiac cell

The cardiac cell or cardiomyocyte is the unit base of the heart. Cardiomyocytes are not all the same in the myocard (e.g. it exists atrial or ventricular cells). However all of them have the same following properties. It consists in an eukaryote tubular cell due to its linear chains named myofibrils, themselves made up of sarcomeres [BDVR+03]. Those sarcomeres are responsible for the mechanical contraction of the cell.

To allow this contraction, the cardiomyocyte needs energy. This energy is provided thanks to the high density of mitochondria through the Krebs cycle [AB21]. However the concentration of ADP/ATP is necessary but not sufficient for the cell contraction. The calcium (mainly present in the sarcoplasmic reticulum), is necessary for the contraction [ECKT17]. Then, a regulation of the calcium concentration in the cell has to be performed between the intracellular space and the extracellular space through the plasmic membrane of the cell (named sarcolemma). This regulation interferes in a series of ionic moves through the sarcolemma to control the balance of the electro-chemical state of the cell.

It exists different kinds of cardiac cells due to its role in the heart. These differences appear (among others) on the cell membrane (kind and number of channels). However, despite these differences, cardiomyocytes basically work in a same way and its electrical activity is closely related to the structure of the sarcolemma.

2.2.1 Sarcolemma

The sarcolemma is a permeable wall (mainly a bilayer of lipids) where some ions and molecules (or macromolecules) can pass through specific channels. These structures

are very useful for the communication between cells but also for the adherence and the molecular transportation for the homoeostasis. The molecular transportation and the ion transportation more precisely are the backbone of the cardiac cycle activity. These transports are specific to each ion and can be active (such as pumps) or passive (such as channels) [AJL+02]. Ionic channels allow a rapid and selective current induced by a stimulation leading to the cardiomyocyte contraction. Then, ionic channels allow a regulation of each ionic concentration from each side of the sarcolemma (given in Table 2.1).

Ion	Intracellular concentration (mM)	Extracellular concentration (mM)
Na^+	~ 20	~ 145
Ca^{2+}	~ 0.0001	~ 2.5
K^+	~ 150	~ 4

Table 2.1: Main intracellular and extracellular ionic concentrations of a cardiomyocyte [Kla11].

Ions being electrically charged, the ionic regulation induces a polarity of the cardiac cells.

2.2.2 Electrical activity

When changes appear in the ionic concentrations (see Table 2.1) from each side of the sarcolemma, the cardiac cell polarity also changes. The permeability of the cell membrane induces a chemical gradient, meaning that sodium and calcium tend to enter into the cell whereas potassium tends to leave the cardiomyocyte (also known as Fick diffusion [Phi06]). However, this chemical gradient is blocked by the electrical equilibrium which prevents ions to pass (in a passive diffusive way) through the sarcolemma.

2.2.2.1 Electro-chemical equilibrium

This electrical equilibrium or Nernst equilibrium [Ste13] is given in Equation (2.1).

$$E_s = \frac{RT}{zF} ln \frac{[s]_e}{[s]_i}, \tag{2.1}$$

where E_s is the electrical equilibrium of species s, R is the universal gas constant ($\approx 8.314 J.mol^{-1}.K^{-1}$), T is the temperature (in Kelvin), z is the valence (ionic charge), F is the Faraday constant ($\approx 96.10^3 C.mol^{-1}$) and $[s]_j$ is the intracellular ($j = i$) or extracellular ($j = e$) concentration of species s. Then, E_s corresponds to the value of the transmembrane voltage such that the ion s does not pass through the membrane (diffusive and electrical forces are counterbalanced).

Changes in the transmembrane potential will then induce changes in the concentrations due to the passive diffusion. Then, to maintain concentrations even after changes in the

transmembrane potential, active membrane structures (pumps or exchangers) allow to maintain the ionic concentrations from each side of the membrane.

2.2.2.2 Stimulation and cardiac action potential

When a cardiac cell is stimulated, the electro-chemical equilibrium is momentarily disturbed, sodium channels are open and an inward current of Na^+ appears. If this electrical stimulation is strong enough, the cell is then depolarised (all-or-none law). The potential difference between the extracellular and intracellular media increases from approximately $-90mV$ to approximately $+20mV$. These values change with respect to the cardiac cell (i.e. atrial, ventricular, ...). To return to the electro-chemical equilibrium, calcium channels are open and an inward current of Ca^{2+} is induced (plateau phase). Finally, to go back to its rest potential ($\sim -90mV$), K^+ leaves the cell through the opened potassium channels (repolarisation phase). These above aspects are quite general and more specific channels exist. The global electrical activity recording (see Section 2.3.1) during a cellular cardiac cycle is named Action Potential (AP). An AP example with its phases is shown in Figure 2.1.

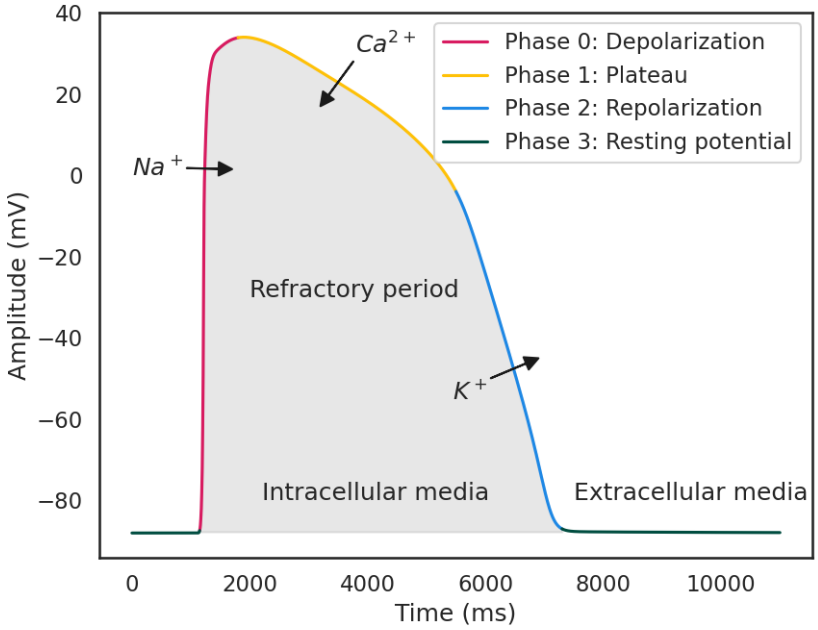

Figure 2.1: Example of action potential with its different electrical phases.

One full AP cycle (as shown in Figure 2.1) corresponds to one cardiomyocyte beat.

2.2.2.3 Propagation

Up to now, we considered an isolated cardiac cell. However, at tissue and heart scale, cells are not exactly beating at the same time. Indeed, once a region starts to beat, the current is transmitted, inducing a contraction of the cell in the neighbourhood. This conduction is allowed by gap junctions [KG96, KS01]. These gap junctions are macromolecular structures composed of two hemi-channels belonging to the sarcolemma of two neighbours cardiomyocytes (see Figure 2.2).

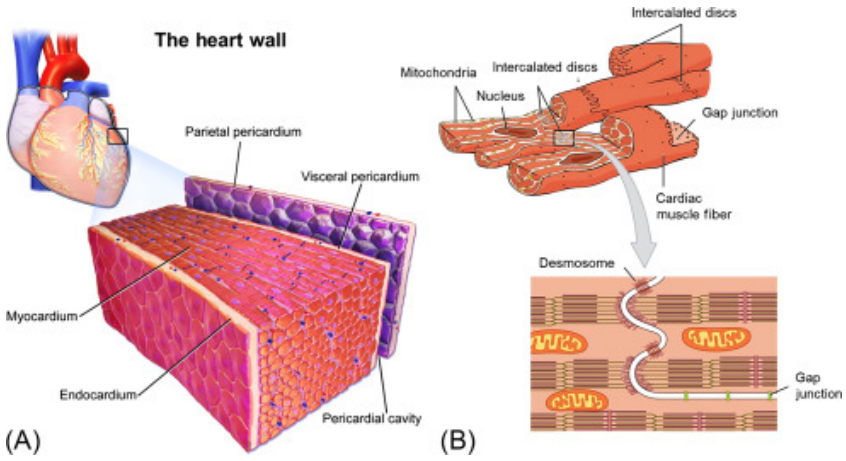

Figure 2.2: Scheme of connected cardiac cells, by [KNG+19] authorized by CCC RightsLink® under license 5253561206879.

They allow the stimulation to the neighbour cardiac cell. To prevent the stimulation of the previous cell (already excited) a refractory period exists. Other way stated, a cardiac cell newly excited has to wait a certain time (refractory period) before being excited once again. This refractory period is induced by the inactivation of the sodium channels. A new contraction is then able once these inactivated channels are reset. At body scale, the usual extracellular recording of the electrical heart activity corresponds to the electrocardiogram (ECG) [Ges89].

2.2.3 hIPSC-CM

In order to perform high throughput screening, pharmacologists need a large bank of cardiomyocytes. However, these cells are difficult to access. Human induced pluripotent stem cell-derived cardiomyocytes (hIPSC-CM) are therefore a good alternative in cardiac safety pharmacology [ASD21, SBM17]. The main idea behind these cells is the potency to reprogram any somatic cell in order to become a pluripotent stem cell. This major discovery was awarded by a Nobel prize to Gurdon and Yamanaka for their works [KC13].

Once these pluripotent stem cells are obtained they can be induced into differentiated cells (e.g. cardiomyocytes: hIPSC-CM) [TTO+07, BKGW12]. The assets of this technique are twofold. On one hand, it avoids using embryonic cells (ethic problems) [Den06, WM03]

and, on the other hand, it allows patient-specific studies [Yam07]. For these reasons, hIPSC-CMs are now widely used in cardiac safety pharmacology for drug assessment and arrhythmia risk studies [TTM+09, KATW15, HAC+13, MdKD+18, KYO+18].

2.3 In vitro electrophysiological devices

To study compound effect on the heart at early stages, *in vitro* assays are performed. These assays are realised on cardiac cells (or cardiomyocytes) which are either isolated or grouped, forming a tissue. These steps are essential in cardiac safety pharmacology to detect arrhythmia risks and/or ionic channel blockade. A compound altering the nominal activity of a cardiomyocyte may impact its contraction and then its electrical activity. This is the reason why electrophysiology is so important in cardiac safety pharmacology. To study this electrical activity, several techniques were developed for compound assessment. In particular, we will focus on Patch-Clamp techniques, a widely used method in electrophysiology. Another promising device described in this manuscript is the Microelectrode Array.

2.3.1 Patch-Clamp techniques

Patch-clamp techniques were originally discovered in the 70's by Erwin Neher & Bert Sakmann [SN84] which granted them to the Nobel prize. These techniques aim at recording electrical activity of a cell from ionic currents passing from the extracellular to the intracellular media. These ionic movements from either side of the cell membrane are characteristic to the cells (e.g. neurons [SSLN10] or cardiac cells [HML+03]). The recorded signal can either be a current or a voltage depending on the protocol used by the analyst. The experimenter can impose a current (current-clamp) or a voltage (voltage-clamp), which leads to procure the other physical quantity using Ohm's law [Kor07]. These recorded signals are essential in safety pharmacology to study compounds and prevent cardiotoxicity and arrhythmia risks. Indeed, the cell membrane (sarcolemma in the case of cardiomyocytes studies) can tell a lot on the cell functionality through the reactions which follow intracellular/extracellular ionic exchanges (see Section 2.2.1). This is one of the main reasons justifying why this technique is widely used in compound investigations in the context of cardiac electrophysiology [KV02, JVD+10, JVM+12].

2.3.1.1 Overview of Patch-Clamp techniques

The principle is based on the glass properties of the pipettes able to stick to the cell membrane. It induces a gigaseal area, implying locally an electrical isolation of the cell membrane ("patch") [CN92]. Patch-clamp techniques are an ensemble of 5 main configurations: Whole cell, Attached cell, Perforated Whole cell, Outside-out and Inside-out which are described in the following section.

Configurations

- Whole cell [FBB02]: This is the most commonly used mode of patch-clamp techniques. In this configuration, the whole cell membrane is involved in the current recording through multiple channels at once. However, the intracellular medium will be replaced by the electrode solution. Action potential models described in Section 3.2 are in a sense similar to this technique. A scheme is given in Figure 2.3.

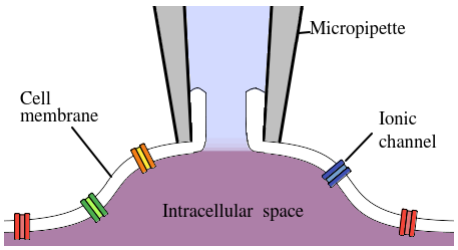

Figure 2.3: Whole cell configuration.

- Attached cell [MSMSR05]: The micropipette (containing a measurement electrode) is in contact with the cell membrane. The suction (forming the patched area) induces a tight seal around it. It follows that measured signal corresponds to the exchange of ions between the micropipette and the patched area. Very few channels (or even a single one) may lie into this patched region. As the membrane is not perforated, the cell is not disturbed, and then mechanisms have a nominal activity. A scheme is given in Figure 2.4.

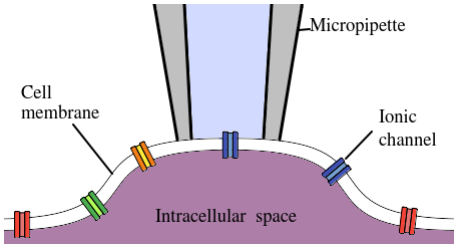

Figure 2.4: Attached-cell configuration.

- Perforated Whole-cell [Lin13]: This configuration has the same target has the Whole-cell case. However, instead of applying a suction, a perforation is realised from the Attached-cell configuration through pore-forming compounds. It allows ions and small molecules to pass through the perforated patch whereas larger molecules and organites cannot pass.

- Outside-out: Once the whole cell configuration is performed, the patched region is detached from the rest of the cell. The selected channel is outside the cell and the detached patch forms a kind of open vesicle attached to the micropipette. It allows to study impact of an element from extracellular medium (micropipette solution) on the selected ionic channel. A scheme is given in Figure 2.5.

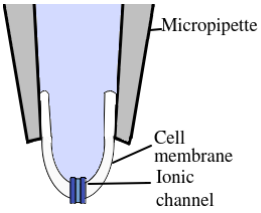

Figure 2.5: Outside-out configuration.

- Inside-out: Here, the patched area is also detached from the rest of the cell. The selection channel is inside the micropipette whereas the cytosolic part of the membrane is exposed to the extracellular medium. This technique is well adapted to study channels activated by intracellular ligands. A scheme is given in Figure 2.6.

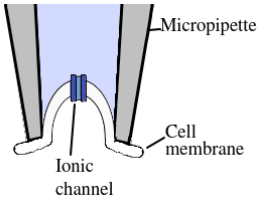

Figure 2.6: Inside-out configuration.

Pros & Cons The existence of these different techniques is justified by several benefits and drawbacks. The experimental protocol has to be driven with respect to the desired study (e.g. ionic channel specific, global activity or intracellular/extracellular point of view). These pros and cons are summarised in Table 2.2.

2.3.1.2 Automated Patch-Clamp

While patch-clamp is essential, various drawbacks exist and have to be overcome. Indeed, this method is very time-consuming and tedious. Thus, automated patch-clamp method was developed in the late 1990's and aims at replacing the manual patch-clamp process to reduce variability and to increase rapidity. Different studies have shown the interest of automated patch-clamp for high-throughput screening in the safety pharmacology context [BF21, YXZ12]. Several kinds of automated patch-clamp techniques exist such as presented in [YXZ12, DBP+08]. One of these techniques is shown in Figure 2.7.

It represents the bottom of one perforated well, on which a cell is attached and perforated. A probe is connected between the extracellular space and the intracellular space allowing the electrical recording of the attached cell.

Configuration	Pros	Cons
Whole-cell	Record the global current. Intra- and extra- cellular components are known.	Cannot measure a unitary current. Dialysis between micropipette medium and intracellular medium.
Perforated Whole-cell	Record the global current. Less damaged cell.	Slow perforation. Tuning cell-dependent.
Attached-cell	Intracellular space preserved.	Unknown transmembrane potential.
Inside/Outside -out	Single channel current measured.	Cannot measure the global current. Cannot study a current if the channel depends on intracellular components.

Table 2.2: Patch-clamp configurations: Pros and Cons.

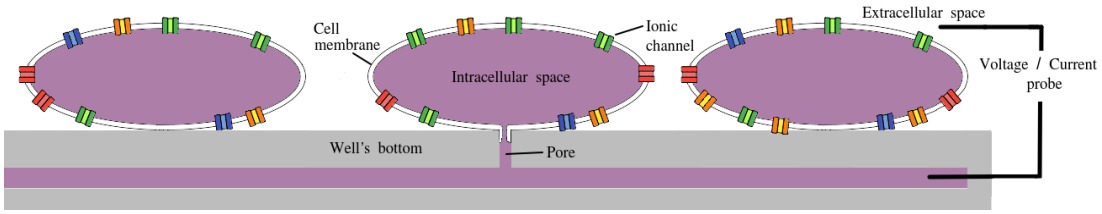

Figure 2.7: Example of automated patch-clamp technique.

2.3.2 Microelectrode Arrays

Microelectrode arrays (MEA) were originally developed in the 50's to measure the neuronal activity [Che07] whereas its first use on cultured cells was performed around twenty years later [TJSL+72, Pin06]. MEA devices correspond to plates on which several electrodes are placed in order to record the electrical cells activities. They were highly developed in the last decades due to its interest in cardiac safety pharmacology [MBGF04, SEG+03, MdKD+18]. These improvements tend to automise the experimental process of electrical activity recordings.

2.3.2.1 Devices

Nowadays, in a high throughput screening context, a MEA is a plate presenting several wells. In each of these wells, some electrodes are placed on its bottom. It exists a wide zoology of MEA with different numbers of wells and electrodes per well. Moreover, in some MEA, electrodes may induce a current in order to diversify the protocols. An example of MEA device is shown in Figure 2.8 and a scheme with a zoom on one well is shown in Figure 2.9.

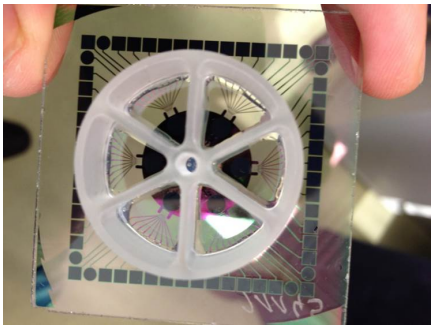

Figure 2.8: Example of MEA device: 6 wells, 9 recording electrodes per well from Multichannel Systems[2].

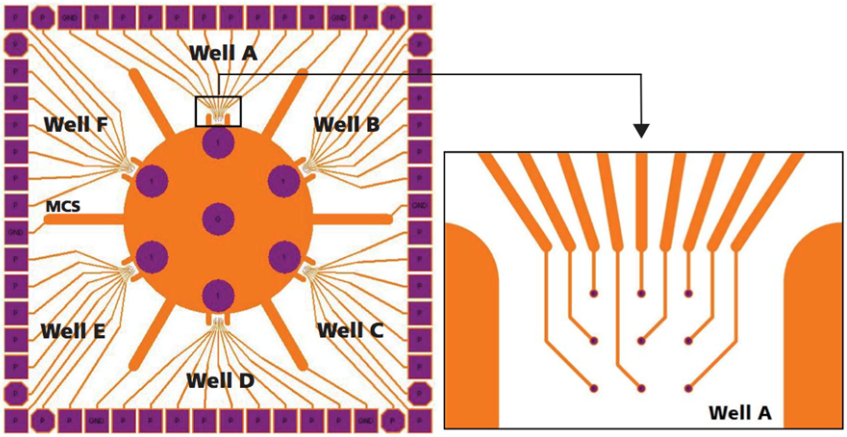

Figure 2.9: Zoom on one well of the above MEA. The four plates at the well border (left panel) are the electrical ground.

To fulfil an experiment, the biologist has to prepare a cell culture and put it inside the wells. At this moment, cells are lining the electrodes. Then, the recording can be performed. It seems quite similar to the automated patch-clamp. However, contrary to this one, cells are preserved (absence of perforation), allowing a non-invasive process. Moreover, the electrical activity being measured at each electrode, recorded signals namely Field Potentials (FP) contain information on the conduction in the tissue. It results in a sense to a kind of electrocardiogram at tissue scale.

2.3.2.2 Extensions

In addition to the electrical activities, several extensions were developed to record other physical quantities. The argument of this is the fact that for some compounds it

[2]Documentation available here.

seems that field potential does not change before and after compound addition (at least to the naked eye), whereas it is known to have an effect. As an example, Blebbistatin, a myosin II inhibitor, is known to reduce the cardiomyocyte contraction while the field potential seems to be the same [LSP15]. The idea is then to measure cells contraction through an impedance signal [OJT+16].

2.4 Conclusion

As the base unit of the heart, the cardiomyocyte is the gateway to preclinical studies and cardiac safety pharmacology more precisely. Its contractility characterisation is induced by an electrical current governed by specific ions passing through the sarcolemma via different macromolecular structures such as channels or pumps. Thus many studies turned to this cardiomyocyte electrical activity, leading to several measurement techniques and tools.

Patch-clamp technique is an ensemble of several configurations depending on the study the experimenter wants to perform: global activity using a current-clamp protocol, specific ionic current activity using a voltage-clamp protocol, ... These techniques are nowadays commonly used in cardiac safety pharmacology and drug development to study exchanges between the extracellular and intracellular medium. In particular, recorded signals such as action potentials are well known and a shape modification induced by a compound can quite easily be identified (sodium, calcium or potassium blockade, early afterdepolarisations, ...).

However, all of these configurations are quite tedious to perform and signal variability may appear. To overcome this, automated patch-clamp seems quite promising. Indeed, it allows a high throughput screening, meaning that at a same time different compounds at a given concentration can be tested.

Another promising technique are MEA devices. They also allow a non-invasive high throughput screening, but at a tissue scale, leading to signals closer to the electrocardiogram.

To preserve the benefit of the high throughput screening, an automated signal analysis has to be investigated to efficiently guide pharmacologists in their future decisions. In complement of these measurement tools, *in silico* models emerged to go deeper in the study of the cardiomycyte functionality.

Chapter 3

Mathematical modelling and simulations

One of the CiPA initiative goals is to integrate experimental data with standardised *in silico* simulations to improve drug assessment [PJYK19, HLLS20]. The use of *in silico* models allows to generate a wide database of simulations, aiming at defining a metric to quantitatively evaluate the impact of a compound on the cardiac electrical activity [PJYK19, Lei20, LRH+19, LMY+20]. These models essentially consist in reproducing Action Potential signals but recent studies have been done on Field Potential signals or other physiological measurements [JCW+20, Küg20]. Their development allows an improvement of the cardiac electrophysiology comprehension at different scales: cell, tissue, heart and body.

Contents

3.1 Introduction

In this manuscript, we will focus on two kinds of electrophysiological signals: Action Potential and Field Potential. While the first is widely used in safety pharmacology, the second one is promising but challenging due to the lack of knowledge. This Simulation part is divided into two sections. The first section (see Section 3.2) is devoted to Action Potential simulations and how these *in silico* models are derived. In the second section (see Section 3.3), we will focus on Field Potential simulations and how to consider the different aspects of the MEA.

3.2 Action Potential simulation

Originally A. Hodgkin and A. Huxley tried to explain the ionic mechanism behind the action potential of a squid axon [HH52] for which the study led to a Nobel prize. The proposed idea was to consider the cell membrane as a dipole. The lipid bilayer part of the membrane is assimilated to a capacitance (as an excellent insulator separating the intracellular and extracellular media) whereas each ion channel carries a current described by the conductance (specific to the considered ion channel) and the voltage by the use of the Ohm's law. An electrical scheme of the model is shown in Figure 3.1.

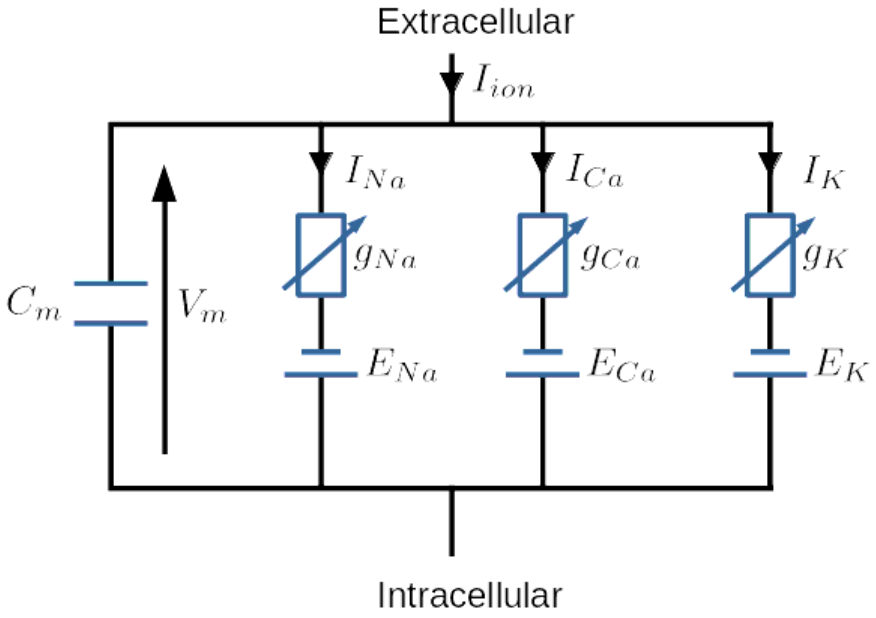

Figure 3.1: Electrical scheme of the Hodgkin and Huxley model considering sodium (Na), calcium (Ca) and potassium (K) channels.

It then results in the model described in Equation (3.1).

$$C_m \frac{dV_m}{dt} = \sum_s \tilde{g}_s(E_s - V_m) + I_{app}, \tag{3.1}$$

where C_m is the membrane capacitance, V_m is the transmembrane potential, t is the time, $\tilde{g}_s$ is the conductance of the channel specific to the ion s, E_s is the reversal potential specific to the ion channel s (see Nernst equilibrium in Section 2.2.2.1) and I_{app} is the stimulation. The first term in the right-hand side of Equation (3.1) can be contracted to I_{ion} as follows:

$$I_{ion} = \sum_s \tilde{g}_s(E_s - V_m). \tag{3.2}$$

However, ion channels contain gates (see Figure 3.2) which can be in an open or closed state, resulting in: $\tilde{g}_s = \tilde{g}_s(t, V_m)$ and can be rewritten as shown in Equation (3.3).

$$\tilde{g}_s = g_s(E_s - V_m) \prod_{i=1}^{Nb_s} \gamma_{s,i}^{p_{s,i}}(t, V_m), \tag{3.3}$$

where g_s is the maximal conductance for the ion channel s, Nb_s is the number of gates for the ion channel s, $\gamma_{s,i}$ is an ODE corresponding to the gate i of the ion channel s activation and $p_{s,i} \in \mathbb{N}^*$ is a constant related to the opening/closing behaviour of the gate.

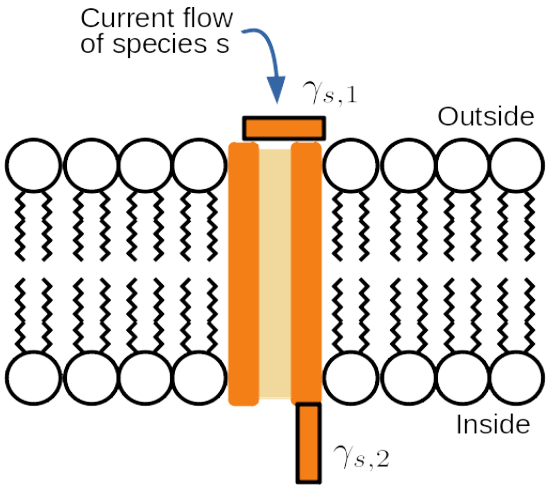

Figure 3.2: Scheme of an ionic channel specific to the ion s which tends to enter into the cell with 2 gates. Gate $\gamma_{s,1}$ is closed and gate $\gamma_{s,2}$ is open. In this configuration, the channel is closed but available for activation. When both gates are open, the channel is activated, whereas when both gates are closed, it is inactivated.

In addition, the ODE depends on parameters specific to each channel and cardiac cell type (i.e. atrial, ventricular, ...). It follows that the general action potential model has the form given in Equation (3.4).

$$C_m \frac{dV_m}{dt} = I_{app} + \sum_s \left(g_s(E_s - V_m) \prod_{i=1}^{Nb_s} \gamma_{s,i}^{p_{s,i}}(t, V_m) \right). \tag{3.4}$$

3.2.1 Different AP models

A wide variety of AP models can be derived from Equation (3.4). Indeed these models can be fitted on atrial cells, ventricular cells, hiPSC-CM (atrial or ventricular) cells, considering concentration dynamics (e.g. calcium transient) or not... Some of these models are given in Table 3.2.1.

Name	Cell type	Number of ODE in I_{ion}
Courtemanche [CRN98]	Atrial human	~ 20
O'Hara Rudy [OVVR11]	Ventricular human	~ 30
Paci [PHASS13]	Atrial/Ventricular hiPSC-CMs	~ 20

Table 3.1: Example of physiological action potential models.

Alternatively, it exists less physiological and more phenomenological models allowing AP simulations. Simpler, these models allow fast simulations but are less accurate. Two of them are described in Table 3.2.1.

Name	Cell type	Number of ODE in I_{ion}
FitzHugh-Nagumo [Fit61, NAY62]	Atrial human	1
Minimal Ventricular [BOCF08]	Ventricular human	3

Table 3.2: Example of phenomenological action potential models.

Among those models, the O'Hara Rudy model (ORd) is one of the most used models in cardiac drug assessment [CFG+16, CJVJS16, PHdK+18]. Indeed, this model is able to reproduce early afterdepolarisations (EAD[1]) which can result in arrhythmia such as torsades de pointes [WGK+10]. Examples of simulated AP are shown in Figure 3.3. The discretisation method used was a backward differentiation formula (BDF3) with a time step $\Delta t = 0.1ms$.

3.2.2 Drug modelling

Two main quantities are considered in cardiac safety pharmacology to study how a compound impacts ionic channels: the compound concentration and the $IC50_s$. The $IC50_s$ is the concentration for which a given compound blocks channels specific to the ion s at 50% of their nominal activity. To render the compound action on the ion channels, the conductance-block model [BPS+06, MCS+11, ZBS+13] is proposed. This model is rewritten in Equation (3.5).

[1]EAD are abnormal depolarisations during phase 2 or 3 of the Action Potential.

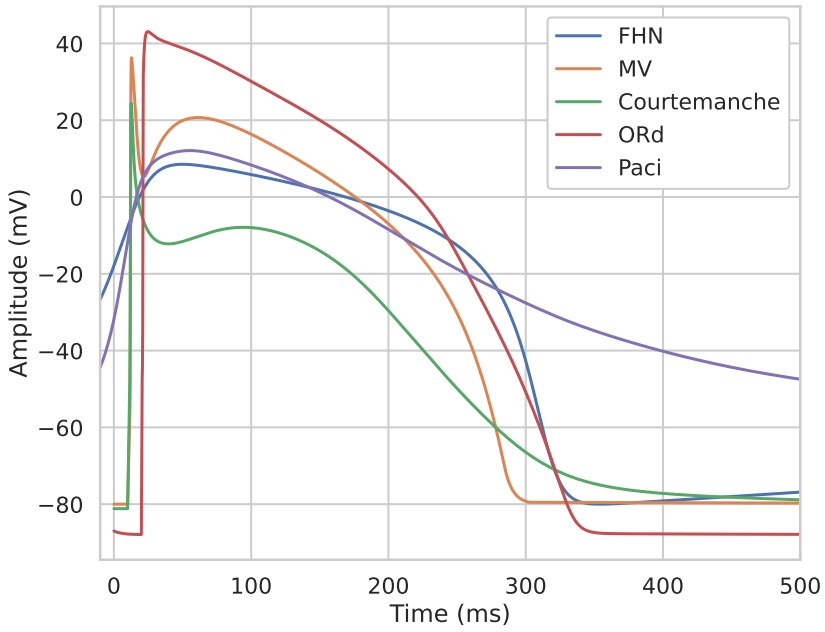

Figure 3.3: Example of AP signals with different models. FHN: FitzHugh-Nagumo.

$$g_s = g_{control,s}\left[1 + \left(\frac{[D]}{IC50_s}\right)^{n_H}\right]^{-1}, \tag{3.5}$$

where $g_{control,s}$ is the conductance of the ion channel s at control case (baseline, corresponding to g_s in Equation (3.3)), $[D]$ is the concentration of the drug and $IC50_s$ is a constant (specific to the compound) corresponding to the concentration at which the ion channel s is blocked at 50%. The parameter n_H is the Hill coefficient, quantifying the interaction between the ligand (compound to simulate) and its binding sites (the ion channel) [Wei97].

Indeed, the model acts as a scale factor between [0,1] in Equation (3.3). When the compound concentration decreases, the scale factor tends to 1 meaning that the ion channel is not affected (control case) whereas an increase of the compound concentration will tend the scale factor to 0 meaning that the ion channel is fully blocked. The scenario where the compound concentration $[D] = IC50_s$ means that the ion channel s is blocked at 50% (as the definition of the $IC50_s$). An example is shown in Figure 3.4 where the corresponding $IC50_s$ are given in Table 3.2.2.

3.3 Field Potential simulation

Field Potentials (FP) are signals recorded by MEA devices (see Section 2.3.2). However, the recorded electrical activity is extracellular (we do not record the transmembrane

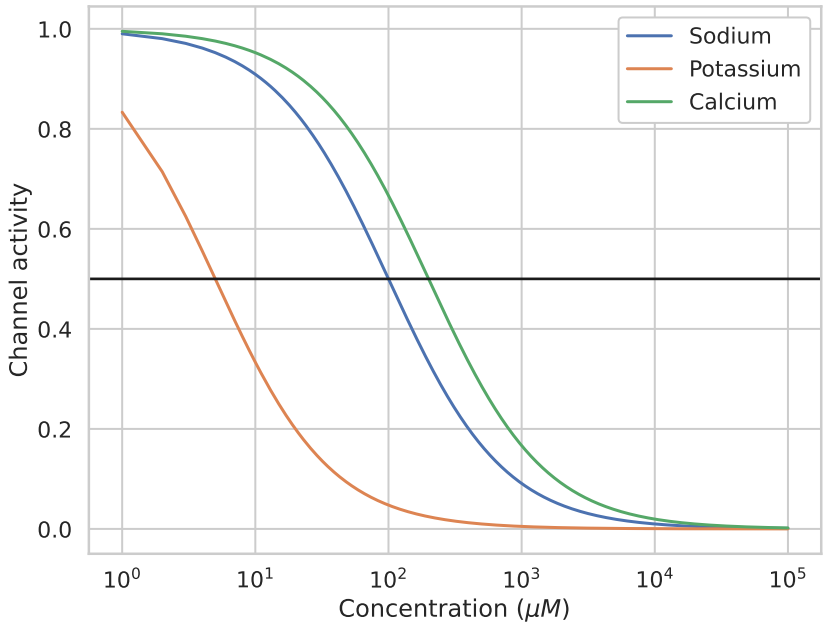

Figure 3.4: Example of dose-response curves. Hill coefficient for each considered ion channel was set to 1. Corresponding $IC50_s$ are given in Table 3.2.2.

Channel	Sodium	Potassium	Calcium
$IC50$ (μM)	100	5	200

Table 3.3: Example of $IC50$ for sodium, potassium and calcium channels used in the conductance-block model to generate Figure 3.4.

potential). Moreover, cardiomyocytes are forming a tissue leading to an electrical interaction. In particular, the conduction is permorfed through the gap junctions (see Section 2.2.2.3).

3.3.1 Finite element mesh

First, we need to physically describe the cell organisation in a MEA well defined as a discretised domain $\mathcal{D} \subset \mathbb{R}^2$. An example of P1 Lagrange finite element mesh is shown in Figure 3.5 (left panel) with the corresponding dimensions (right panel).

3.3.2 Bidomain model

Bidomain equations describe the electrical propagation in a cardiac tissue (e.g. represented by the finite element mesh described in the previous section). They were formulated in the late 70's [Tun78a] and aim at studying the electrical activity at heart scale and

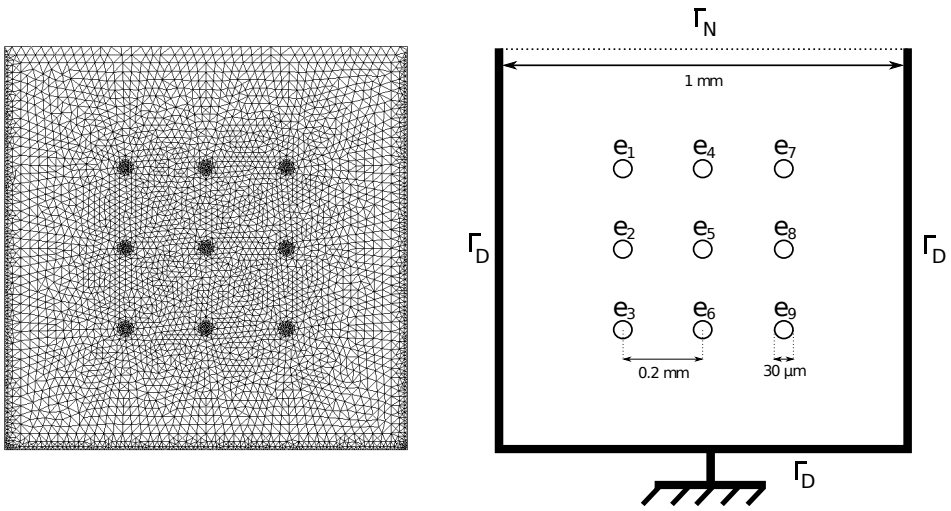

Figure 3.5: Example of P1 Lagrange finite element mesh (left panel), with the corresponding physical dimensions (right panel). Other descriptions of this MEA are given in Figures 2.8-2.9 of Section 2.3.2.1.

body scale with ECG simulations [PBC05, ZBS+13]. Its derivation is based on the cable theory and considers the whole tissue (or heart) characterised by an intracellular and extracellular space separated by a membrane [GM83].

$$\begin{cases} A_m\Big(C_m \frac{\partial V_m}{\partial t} + I_{ion}(V_m, \gamma)\Big) - \nabla \cdot (\sigma_i \nabla V_m) - \nabla \cdot (\sigma_i \nabla \phi_e) = A_m I_{app} \\ -\nabla \cdot \Big((\sigma_i + \sigma_e)\nabla \phi_e\Big) - \nabla \cdot (\sigma_i \nabla V_m) = 0 \end{cases}, \tag{3.6}$$

where A_m is the ratio of membrane area per unit volume, C_m is the membrane capacitance, V_m the transmembrane potential, I_{ion} is the current given by the action potential model (see Equation (3.2)), σ_i and σ_e the intracellular and extracellular conductivity respectively, ϕ_e is the extracellular potential and I_{app} the stimulation.

In the MEA case, we consider a modified version of the bidomain equations (see Equation (3.7)) to take into account the interaction with the electrodes.

$$\begin{cases} A_m\Big(C_m \frac{\partial V_m}{\partial t} + I_{ion}(V_m, \gamma)\Big) - \nabla \cdot (\sigma_i \nabla V_m) - \nabla \cdot (\sigma_i \nabla \phi_e) = A_m I_{app} \\ -\nabla \cdot \Big((\sigma_i + \sigma_e)\nabla \phi_e\Big) - \nabla \cdot (\sigma_i \nabla V_m) = \frac{1}{z_{thick}} \sum_{e_k} \frac{I_{el}^k}{|e_k|} \mathbb{1}_{e_k} \end{cases}, \tag{3.7}$$

where $|e_k|$ denotes the surface of the electrode k, $\mathbb{1}_{e_k}$ denotes its indicator function and z_{thick} is the thickness of the cell layer. The right-hand side term of the second equation, stands for the electric current I_{el}^k going through the electrode located at e_k. More details are given in [ABC+18]. The I_{el}^k current is computed through an imperfect model of the electrode (see Section 3.3.3).

3.3.2.1 Boundary conditions

Let $\mathbf{n}$ be the outward normal to the boundary of the domain $\mathcal{D}$ (i.e. a well). The bidomain model extended to the MEA (see Equation (3.7)) is completed by the boundary conditions defined in Equation (3.8).

$$\begin{cases} \sigma_i \nabla \phi_i \cdot \mathbf{n} = 0 \\ \begin{cases} \phi_e = 0 \text{ on the region connected to the ground} \\ \sigma_e \nabla \phi_e \cdot \mathbf{n} = 0 \text{ elsewhere} \end{cases} \end{cases}, \tag{3.8}$$

where $\phi_i = V_m + \phi_e$ is the intracellular potential. An example is given in the right panel of Figure 3.5, where Γ_N stands for the Neumann boundary condition and Γ_D stands for the Dirichlet boundary condition.

3.3.2.2 Source term: I_{app}

The current $I_{\text{app}} = I_{\text{app}}(x, y, t)$ is the origin of the activation. The source is supposed to be located in a unique region and is defined as follows:

$$I_{\text{app}}(x, y, t) = \begin{cases} I_0 \exp\left[\frac{(t-t_0)^2}{2\sigma^2}\right] & \text{if } (x - x_0)^2 + (y - y_0)^2 \leq r^2, \\ 0, \text{ otherwise,} \end{cases} \tag{3.9}$$

where the position (x_0,y_0) is drawn randomly. Parameter r is the radius of the source, I_0 is the maximum stimulation value and t_0 is the time when I_{app} is at its maximum.

Remark 1
Sometimes, it may have multiple localised sources in a same well. However, based on in vitro data, it seems that a unique source appear most of the time (as obtained through activation maps in [INNW+17, LBM+17, ZSS+17]).

3.3.3 Electrode model

To take the impact of the electrodes on the signal into account, an imperfect electrode model [RBZ+17] is coupled with the bidomain equations. The model is described in Equation (3.10).

$$\frac{dI_{el}^k}{dt} + \frac{I_{el}^k}{\tau} = \frac{C_{el}}{\tau} \frac{d}{dt} \phi_{e,mean}^k, \tag{3.10}$$

where $\phi_{e,mean}^k$ is the averaged extracellular potential on the electrode e_k, R_{el} and C_{el} are the electrode resistance and electrode capacitance respectively and R_i is the internal resistance of the measurement device. $\tau = C_{el}(R_i + R_{el})$ is the time constant of the RC circuit. Then, the field potential ϕ_f^k measured on the electrode e_k is given by $\phi_f^k = R_i I_{el}^k$. The electrical scheme of the imperfect electrode model is shown in Figure 3.6.

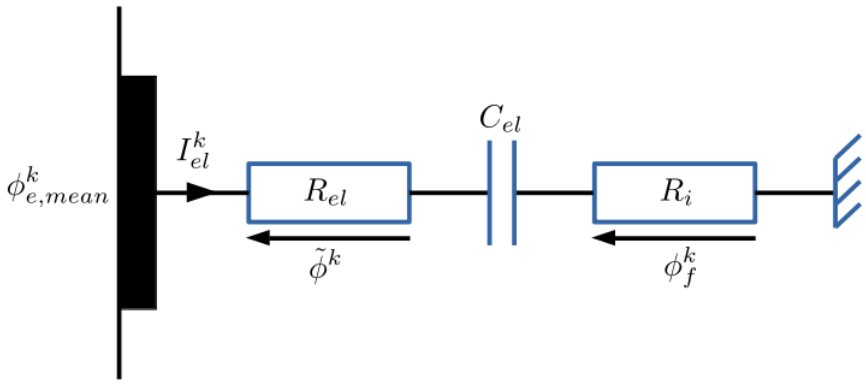

Figure 3.6: Electrical scheme of the imperfect electrode model.

3.3.4 Heterogeneity

Although cultured hiPSC-CMs are highly pure [MGF+11, FYK+20], some of them may not beat, or present another phenotype: atrial instead of ventricular as an example. To simulate this heterogeneity in the cell distribution inside a well, a space stochastic process was introduced, similarly to what was proposed in [TRLG18]: let $(Z, \mathcal{A}, \mathbb{P})$ be a complete probability space, Z being the set of outcomes, $\mathcal{A}$ a σ-algebra and $\mathbb{P}$ a probability measure:

$$c(x,\zeta) : \Omega \times Z \to [0,1]. \tag{3.11}$$

A hypothesis on the correlation of the process was made and expressed in Equation (3.12):

$$f_c\left[\begin{pmatrix}x\\y\end{pmatrix}, \begin{pmatrix}x'\\y'\end{pmatrix}\right] = \exp\left[-\frac{(x-x')^2+(y-y')^2}{2l_c^2}\right], \tag{3.12}$$

that is, the correlation is normal and its length l_c was set to 0.25 mm, which corresponds approximately to the distance between two electrodes. A Karhunen-Loève expansion based on the diagonalisation of the correlation kernel was used in order to generate the heterogeneity fields (see [TRLG18] for details). As shown in [BGMW19] in a case of a low value for l_c, the medium is homogenised which leads to a decrease of the repolarisation phase. The parameter $l_c = 0.25$ mm was a good choice to qualitatively reproduce FP signals and more precisely the repolarisation phase.

3.3.5 Example of applications

3.3.5.1 Field Potential simulation at control case

An example of Field Potential simulations with parameters given in Table 3.4 is presented here.

A_m	C_m	σ_e	σ_e
$200.0cm^{-1}$	$1.0\mu F.cm^{-2}$	$0.002\mu S.cm^{-2}$	$0.002\mu S.cm^{-2}$

Table 3.4: Bidomain equations parameters used for the example.

The time step is $0.1ms$. Parameters used for the imperfect electrode model are given in Table 13.2 and boundary conditions are given in the right panel of Figure 3.5. Parameters of the stimulation are $50\mu m$, $t_0 = 5ms$, $I_0 = -80pA/pF$ and $\Delta t = 4ms$.

C_{el}	R_i	R_{el}
$1nF$	$2M\Omega$	$10M\Omega$

Table 3.5: Parameters used for the imperfect electrode model.

The heterogeneity field is given by the left panel of Figure 3.7. The action potential used is the MV model, considering epicardial cells and endocardial cells (parameters given in [BOCF08]) for the heterogeneity. Right panel of Figure 3.7 shows the propagation wave at time $t = 8ms$.

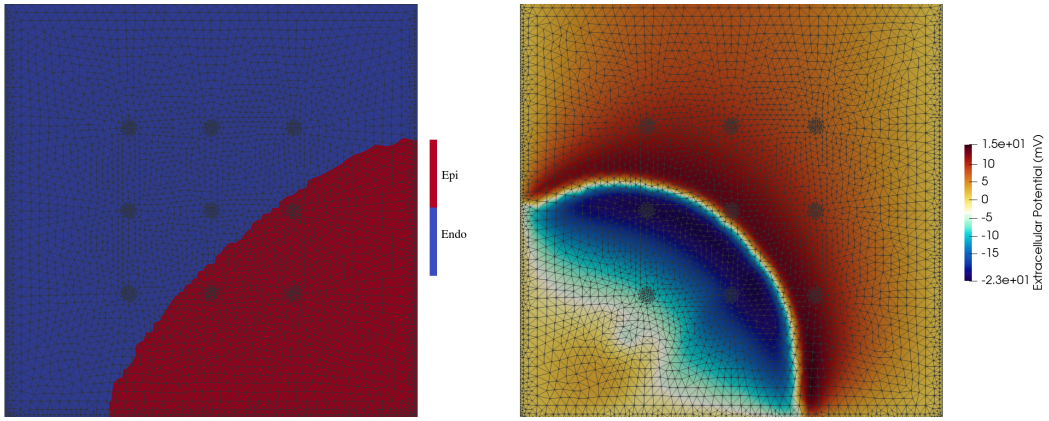

Figure 3.7: Example of simulation. Left: Heterogeneity field (using Epi- and Endo- cardial cells parameters of the MV action potential model). Right: Extracellular potential on the whole well at $t = 8ms$.

Action Potentials and Field Potentials at the electrodes are shown in Figure 3.8.

3.3.5.2 Example of EAD simulation

Using the conductance-block model, we can simulate known or unknown compounds (for instance, by randomly blocking sodium, calcium and/or potassium channels). Figure 3.9 shows an example of simulated early afterdepolarisation (EAD) using the ORd model, as a result of blocking IKr current at 93.5%.

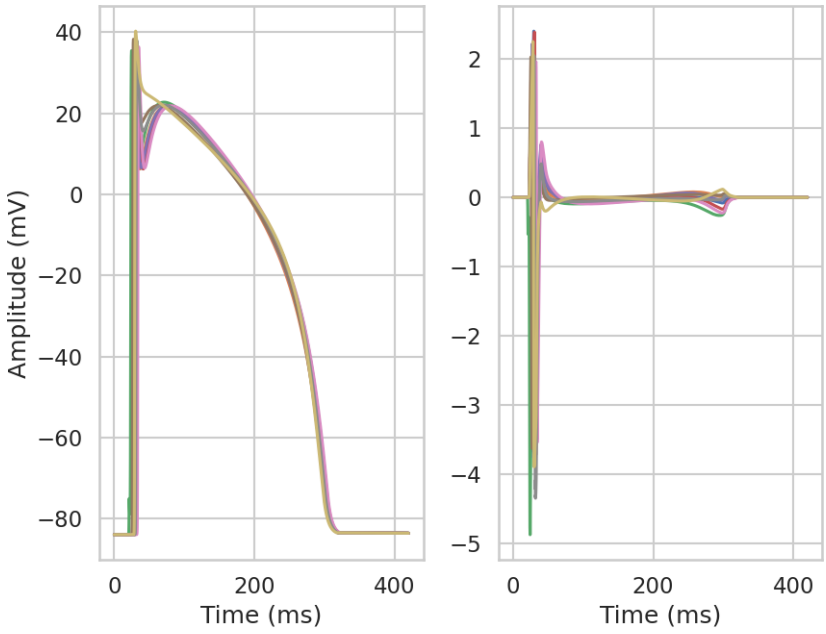

Figure 3.8: Example of simulation. Left: Action Potentials at the electrodes. Right: Field Potentials at the electrodes.

The FP and intracellular calcium transient shapes are in good qualitative agreement with *in vitro* experimental signals [NMM+14, CWL+15, YKI+18].

3.4 Conclusion

While the first *in silico* Action Potential model has been derived in the 50's, this field of research is still in progress for several reasons such as the consideration of new cells (e.g. hiPSC-CM) or the consideration of intracellular organites (e.g. sarcoplasmic reticulum). A wide variety of models exists and can be categorised according to two main types: phenomenological (e.g. FHN or MV) or physiological (e.g. Courtemanche, ORd or Paci). The drug modelling can be achieved by combining an Action Potential model with a conductance-block model which takes into account two pharmacological quantities: the compound concentration and the $IC50$ for considered ion channels.

Action Potential models are essential to mimic the cardiac electrical activity at higher scales such as tissue or body. They are coupled with bidomain equations, governing the electrical propagation in a tissue. The Field Potential simulation can be achieved by resolving these equations on a finite element mesh of the well with suitable boundary conditions. Moreover, different Action Potential models or different parameterisations of an Action Potential model can be used to introduce some heterogeneity in the well.

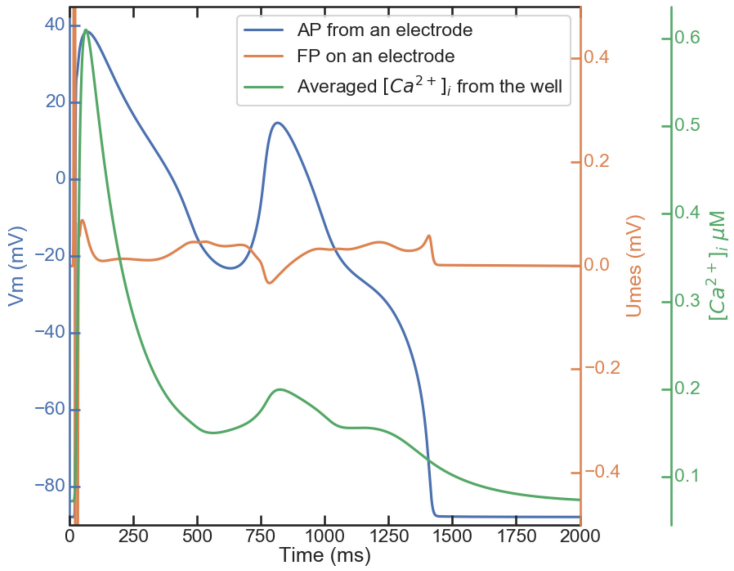

Figure 3.9: EAD simulation. Transmembrane action potential (AP, blue), extracellular field potential (FP, orange) recorded at one electrode and intracellular calcium transient trace (green) in a simulated EAD case.

The ORd Action Potential model combined with bidomain equations leads to qualitatively good signals in EAD case if compared with *in vitro* experimental traces.

Some improvements could however be done by considering the coating present at the bottom of the well. Indeed, to attach cells to the bottom of the well, a coating of fibronectin or collagen is performed in a first step [Bla13].

Chapter 4
Summary

Electrophysiology is the entry point of the cardiac safety pharmacology resulting into two main branches at early stages of drug development: *in vitro* and *in silico* assays.

In vitro experiments have been made possible through the development of measurement tools and techniques, Patch-Clamp being the most used. On the other hand, the access to cardiac cells was improved by properties of cells to be differentiated into cardiomyocytes. More recent measurement tools have been developed to allow high throughput screening such as automated patch-clamp and MEA.

In silico models were performed to go deeper into the investigation of cardiac cells functionalities. Based on the original Hodgkin and Huxley model, they were extended and fitted by considering specific cardiac cells and/or concentration dynamics. Conductance-block model allows the simulation of compound effects, considering its concentration and the corresponding $IC50_s$. The derivation of the bidomain equations based on the cable theory aims at describing the electrical propagation into a tissue. Coupled with an action potential model, it allows to simulate FP signals from a finite element mesh of the well and suitable boundary conditions.

The complementarity of these two branches has been highlighted in various works [RCAG+15, Küg20] and confirms the objectives of the consortiums (CiPA, JiCSA, ...) for new guidelines. The access to an *in vitro* and/or *in silico* database tends to consider probabilistic and statistical approaches to improve pharmacologists decisions. It then, opens the door to the use of machine learning techniques.

Part II

Methodology

Introduction

Computer performances as well as the amount of available data allow nowadays to help scientists to go further in their investigations. Indeed machine learning techniques aim at automatically treating data based on a probabilistic and statistical point of view. However, it remains some domains in which there is a lack of data. This is particularly the case in cardiac safety pharmacology, either because the technique is too slow (e.g. patch-clamp), or it is too recent (e.g. automated patch-clamp and MEA). Moreover this data contain a high quantity of information (being time series). These two points lead to a so-called high-dimensional/low sample size regime (introduced in the Introduction). To overcome this, dimension reduction methods were developed. However, these techniques are not necessarily well adapted in classification task context. Indeed, the method can either be too simple (e.g. Feature Selection: FS), not answering the classification task (e.g. Principal Component Analysis: PCA), not sparse (e.g. Partial Least Squares: PLS) or depending on many parameters (e.g. Neural Networks: NN). The first part of this chapter presents a sparse goal-oriented dimension reduction method respecting the following constraints:

- Sparse dimension reduction,
- Oriented to maximise the classification success rate,
- Few parameters to tune,

in respect with the problematic described in the Introduction.

On top of that, available data may not cover the whole observable space, some of them might be too noisy to be considered or even wrongly labelled in the case of classification task. To struggle with this, numerical simulations may help if a mathematical model on the problem we are working on exists. To deal with these particular cases, many fields of study arose such as augmented set construction or instance selection. For the same drawbacks described in the dimension reduction methods existing, the second part of this chapter is devoted to a method aiming at constructing an augmented set maximising the classification success rate under the following constraints:

- Fewer samples as possible (to reduce training time).
- Oriented to maximise the classification success rate.
- Few parameters to tune.

The two following chapters of this part led to two submitted papers.

Chapter 5

Double Greedy Dimension Reduction method

In numerous classification problems, the number of available samples to be used in order to construct a classifier is small, and each sample is a vector whose dimension is large (with respect to the number of samples). This regime, called high-dimensional/low sample size is particularly challenging when classification tasks have to be performed. To overcome this shortcoming, several dimension reduction methods were proposed. This work investigates a greedy optimisation method that builds a low dimensional classifier input. Some numerical examples are proposed to illustrate the performances of the method and compare it to other dimension reduction strategies.

Contents

5.1 Introduction

This work investigates a method of dimension reduction applied to classification problems. These arise in many areas of applied sciences in which data are queried to provide predictions in a form of yes/no answers or more elaborated classification outcomes. Often, prior of classification, data is pre-processed in order to train in a more effective way a classifier. Part of the pre-processing phase takes the form of a linear or non-linear dimension reduction. Hereafter we propose a systematic way of performing this task.

Let $\mathcal{G}$ be an ensemble of signals, provided from experimental measurements, numerical simulations (or both). Let $n_s \in \mathbb{N}^*$ be the number of samples that will be used to train the classifier: for each $\mathcal{G}^{(i)}, i = 1, \dots, n_s$ a set of $n_g \in \mathbb{N}^*$ quantities are extracted from the signal. These can be either informed linear or non-linear forms identified by experimental insight or more agnostic features, such as point values of the signal, local average, Fourier or Wavelets coefficients. We refer to the set of these quantities for all the available signals as the dictionary entries $G_j^{(i)} \in \mathbb{R}, i = 1, \dots, n_s, j = 1, \dots, n_g$. The present work deals with classification problems, namely, given an observable signal coming from a physical system, we want to determine to which class in a set of possible classes the system belongs to.

In the present work, for sake of simplicity, the method is derived in the case of binary classification: its extension to multiple classes is straightforward. The methodology presented is general, and it was motivated by classification problems arising in biomedical engineering, in which the problems at hand can sometimes be in a different regime with respect to the ones classically addressed in Machine Learning. Indeed, as in other fields of science and engineering, the size n_g of quantities that can be extracted from the signal can be extremely large. Moreover, the number of available samples n_s , due to experimental constraints and to the complexity of the systems at hand, can be small if compared to n_g. This regime, called *high dimensional/low sample size* in the learning community is particularly critical when performing classification and regression tasks. The mathematical reason is that we wish to identify a function whose domain dimension n_g is large, and hence we are exposed to the phenomenon of the curse of dimensionality, introduced for the first time by Bellman in [Bel15] and related to learning theory in [SZ03].

In [CDD+12, FSV12, MUV15] a theoretical analysis is proposed that describes the ability of approximating a high-dimensional ridge function by point queries and how the curse of dimensionality can be eventually circumvented. From a probabilistic viewpoint, for a given sample size, when the dictionary size becomes too large, the classification error increases: this is referred to as Hughes phenomenon [Hug68, Tru79]. This regime is appearing in various areas of science and it is nowadays widely studied [DCZ+13, HMN05, LHNM08, Mec12].

To overcome this difficulty, several strategies have been devised in the literature, involving dimension reduction and sparsification. There is a vast literature on the setup of sparse classifiers. This is often obtained by an optimisation problem involving the ℓ^1 distance. For an extensive overview of these works the reader is referred to [HTW19] and to the seminal works [DET05, CRT06, BMB08]. Recent works have been proposed, as for

instance [CF20, DHM20]. Among the works done in this field, the most similar in spirit to the work proposed here is [ZRTH03], in which an ℓ^1 optimisation is used to construct a sparse input space of a Support Vector Machine classifier. With respect to this work, there are several differences: in the present work we try to setup a goal oriented dimension reduction which aims at improving a classification score, but which is independent of the classifier chosen. Moreover, we adopt a greedy strategy in order to promote sparsity.

Concerning the dimension reduction of the input space (consider [Fod02] for an overview), this was considered in machine learning applications in [GF15, HJP03, KPZ07, LWZY17]. In most of the references, a dimension reduction strategy is applied and the results in terms of classification are then analysed.

Several methods were proposed, for instance, in metric learning [BHS13], which have an analogous goal and are similar, in the spirit, to what is proposed in the current work. Among the methods, we cite and comment the ones which are related the most, and perform some numerical tests. In partial least square (the reader can refer to [RK05] for an overview), the objective is to maximise the correlation between two given sets of variables. This can be adapted to classification and represents a viable way to perform dimension reduction. The main difference with respect to the present work is that we do not attempt to maximise correlation; instead, the method maximises a classification score, which is discussed later on. In the Average Neighbourhood Margin Maximisation (ANMM) [WZ07], the authors propose to construct a linear subspace (to which data are projected on), by pulling together, in a neighbourhood, points with the same label, and try to maximise the distance of points with different labels. This method is similar to what is proposed in the present work, with some differences: the score which is maximised here is not related to distance or margin, per se, it is based on a probabilistic argument and it is the measure of the success event related to the classification; moreover, the linear subspace is found in a greedy parsimonious way, to promote, as much as possible, sparsity. In neighbourhood component analysis (NCA), described in [BHS13, QSHZ15], a linear embedding is learned from data, which, at once, performs dimension reduction and metric learning. It is meant to maximise the classification score of KNN classifiers. In the present work we do not seek to learn a metric, but to maximise a measure. The resulting method could improve the classification score of all kinds of classifiers, as the numerical experiments show.

The proposed method consists in projecting the dictionary entries into a low-dimensional linear subspace (obtained by a sparse linear combination of the entries), which is computed in order to optimise the classification success rate. From a dimension reduction point of view, the method proposed can be considered as a goal-oriented dimension reduction.

5.1.1 Notations and assumptions

Let $X_{n_g} \in \mathbb{R}^{n_g}$ be a random vector of the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We assume that the probability function (pdf) of X_{n_g} is a mixture of the form:

$$\rho(g) = \pi_0 \rho_0(g) + \pi_1 \rho_1(g), \tag{5.1}$$

where $\rho_i(g) = \rho_i(g|y_* = i)$ is the conditional probability of g given that its label is $y_* = i$. The scalars π_i are the weights of the mixture and they can be seen as the *a priori* probability mass of being in the class i. It holds $\pi_0 + \pi_1 = 1$. A classifier is defined in Definition 4 with $n = n_g$.

Definition 4
Let $g \in \mathbb{R}^n$ be an observation, paired with a label y. A binary classifier is a function $\mathcal{C}_n$ such that the following holds:

$$\begin{array}{l} \mathcal{C}_n : \mathbb{R}^{n_g} \to \{0,1\} \\ g \to y \end{array} .$$

Some geometrical notations are introduced. Let $k \leq n_g$. The Grassmann manifold Gr_{k,n_g} is the set of k-dimensional linear subspace of R^{n_g}. The method proposed in the present work can be seen as an optimisation on the compact Stiefel manifold, denoted by $\mathcal{M}_{k,n_g}$, whose definition is recalled in Definition 5. An element of the Stiefel manifold will be denoted by M.

Definition 5
A real Stiefel manifold $\mathcal{M}_{k,n_g}$ is a set of all the k-frames in $\mathbb{R}^{n_g}$:

$$\mathcal{M}_{k,n_g} \stackrel{\Delta}{=} \{Y = (Y_1, \ldots, Y_k), Y_i \in \mathbb{R}^{n_g} \mid Y_i^T Y_j = \delta_{ij}, \forall 1 \leq i,j \leq k\},$$

so that the elements of the compact Stiefel manifold are the matrices of $M \in \mathbb{R}^{n_g \times k}$ with orthonormal columns.

The Stiefel manifold $\mathcal{M}_{n_g,n_g} = \mathcal{O}(n_g)$ is the orthogonal group. An element $R \in \mathcal{O}(n_g)$ satisfies $R^T R = RR^T = I_{n_g}$. An element of $\mathcal{M}_{n_g,n_g}$, can be seen, roughly speaking, as the concatenation of an element of the Stiefel manifold, and an element belonging to the orthogonal complement of the subspace spanned by the columns of M, denoted $M^\perp$, whose columns are orthonormal: $R = [M, M^\perp]$. Let us consider the endomorphism induced by R, and how the probability ρ is transformed accordingly. The change of coordinates $g = R\xi$ is applied to the expression in Equation (6.1) leading to:

$$\rho(\xi) = \pi_0 \rho_0(R\xi) + \pi_1 \rho_1(R\xi),$$

that holds since $det(R) = 1$.

5.2 Method

The method is detailed. An element M of a Stiefel manifold is used to reduce the input dimension: $x \in \mathbb{R}^k$ (the dimension k is, also, an outcome of the proposed method). Let $\mathcal{C}_k$ be a classifier in the projected space of dimension k (see Section 5.1.1 and Definition 4 with $n = k$). It is defined as:

Definition 6
The classifier $\mathcal{C}_k$ in the subspace of dimension $k \ll n_g$ is defined as follows:

$$\begin{array}{l} \mathcal{C}_k : \mathbb{R}^k \to \{0,1\} \\ x = M^T g \to y \end{array},$$

where $g \in \mathbb{R}^{n_g}$ is an observation, $M \in \mathcal{M}_{k,n_g}$ and y is the label in the projected space.

The objective is to find $M \in \mathcal{M}_{k,n_g}$ which maximises the success rate of the classifier $\mathcal{C}_k$. In particular, we will introduce an objective function related to the classification success rate, intrinsically related to the ability of distinguishing the elements belonging to two classes. As a consequence, the proposed method applies to all kinds of classifiers.

5.2.1 Classification score in the reduced space

The classification score is investigated and its relation to the score in the dictionary space is derived. Roughly speaking, reducing the dimension also reduces the amount of information the input carries about the classification output. This loss has to be quantified and minimised.

First, a consideration on the projected density on a Stiefel manifold element is presented, which will be used to derive the relationship between the classification score and the total variation in the reduced input space.

By the properties of orthogonality of the elements Stiefel manifold and one element belonging to the space orthogonal to the space spanned by its columns, the pdf p in the projected space of dimension $k < n_g$ corresponds to the marginals of ρ as shown in Equation (5.2). Indeed, let $M \in \mathcal{M}_{k,n_g}$ and $R = [M, M^\perp]$. Let an input $x = M^T g$; we denote by $\xi \in \mathbb{R}^{n_g}$ the vector $\xi = R^T g$. It follows that $x = [\xi_1; \ldots; \xi_k]$. Since R is an element of the orthogonal group, it holds:

$$p(x) = \int_{\mathbb{R}^{n_g - k}} \rho(\xi) d\xi_{k+1} \ldots d\xi_{n_g}, \tag{5.2}$$

and hence:

$$p(x) = \pi_0 p_0(x) + \pi_1 p_1(x). \tag{5.3}$$

An important consequence is that p is a mixture of the same form as ρ, and, moreover, $p_i(x)$ is the conditional probability density of x given that its label $y_* = i$:

$$p_i(x) = p(x|y_* = i), \tag{5.4}$$

for $i = 0$ or 1.

Remark 2
The element $M^\perp$ is arbitrary, and defined up to a unitary transformation. Remark, however, that the result on the transformation of the probability density $p(x)$ does not change by virtue of the operation of marginalisation.

The input space is subdivided into three distinct regions, in relation to what the classifier $\mathcal{C}_k$ (see Definition 6) would provide, based on a probability argument. We denote by $S_0 \subseteq \mathbb{R}^k$, $S_1 \subseteq \mathbb{R}^k$ and $S_2 \subseteq \mathbb{R}^k$:

Definition 7

$$\begin{cases} S_0 \triangleq \{x = M^T g \in \mathbb{R}^k \mid \pi_0 p_0(x) > \pi_1 p_1(x)\} \\ S_1 \triangleq \{x = M^T g \in \mathbb{R}^k \mid \pi_0 p_0(x) < \pi_1 p_1(x)\} \\ S_2 \triangleq \{x = M^T g \in \mathbb{R}^k \mid \pi_0 p_0(x) = \pi_1 p_1(x)\} \end{cases}.$$

It follows that:

- $S_i \cap S_j = \emptyset, \forall i \neq j$.
- $\cup_{i=0}^{2} S_i = S \subseteq \mathbb{R}^k$.

Let (g, y_*) be a pair such that $g \in \mathbb{R}^{n_g}$ is an observation and $y_* \in \{0, 1\}$ the corresponding label (the true label). Let A_S be the ensemble of the success events, that is when the classifier $\mathcal{C}_k$ provides as result $y = y_*$. The set of success events can be defined as:

Definition 8

$$\begin{cases} A_{S_0} \triangleq \{y_*, x = M^T g \mid (y_* = 0) \wedge x \in S_0\} \\ A_{S_1} \triangleq \{y_*, x = M^T g \mid (y_* = 1) \wedge x \in S_1\} \\ A_{S_2} \triangleq \{y_*, x = M^T g \mid (y_* = 0{,}1) \wedge x \in S_2\} \end{cases}.$$

And,

$$A_S \triangleq \cup_{i=0}^{2} A_{S_i}.$$

Remark that the sets A_{S_i} define the cases in which the label of the classification obtained by the classifier ($x \in S_0$ would be classified $y = 0$) corresponds to the true labels. The success rate is henceforth related to the measure of these sets:

$$\mu(A_S) = \int_{S_0} \pi_0 p_0(x) dx + \int_{S_1} \pi_1 p_1(x) dx + \frac{1}{2} \int_{S_2} p(x) dx. \tag{5.5}$$

The $\frac{1}{2}$ factor is justified by the fact that we expect to have half of the realisations to be well classified on S_2. This score is analogous to the excess risk measure proposed in [BCDD14] which consists in evaluating a regression function over the symmetric difference between the true sets (of each class) and the sets (of each class) obtained through a Bayesian classifier in the context of set estimation.

One of the contributions of the present work is to relate this measure of the success rate for a classification problems to other measures which are widely used in the literature.

An important aspect is that the proposed classification assessment score also applies to data distributions with very mild regularity assumptions. This is presented in the following sections.

5.2.1.1 Relation to the total variation

In order to quantify the success rate of the classification, distances or divergences between densities are commonly used. We denote by δ_{TV} the total variation [BGvdM92, NP16] (see Definition 9[1]). The total variation is a f-divergence [Csi64] which is also a metric over the probability densities.

Definition 9
Let P and Q be two probability distributions on $(\Omega, \mathcal{A})$ (with Ω the sample space and $\mathcal{A}$ a σ-algebra) and p and q the corresponding pdf. Then, the total variation is:

$$\delta_{TV}(P,Q) = \frac{1}{2}\int_{\Omega} |p(x) - q(x)|dx.$$

The pertinence of the total variation in relation to classification can be hinted by the following consideration. When the total variation is 0, the probability distributions corresponding to the two classes coincide almost everywhere. It means that for any observation (up to a zero measure set), we could attribute either 0 or 1 and no discrimination between the two classes would be possible based on a probability argument.

In the following of this paper, we make the hypothesis that the total variation between ρ_0 and ρ_1 is strictly positive, that is $\delta_{TV}(\rho_0, \rho_1) > 0$. In the case of binary classification, we also assume that $min(\pi_0, \pi_1) > 0$. We show hereafter that the measure of success presented in Equation (5.5) is related to the total variation between the densities p_0 and p_1:

Proposition 1
Let $p(x)$ be defined as in Equation (5.3)-(5.4), and the quantity $\mu(A_S)$ be defined as in Equation (5.5). It holds:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big(\int_S |\pi_0 p_0 - \pi_1 p_1|dx\Big). \tag{5.6}$$

The demonstration of Proposition 1 is given in Section 5.5 in the Appendix.

Remark 3
From Equation (5.6) obtained in the proof, we can directly see that $\frac{1}{2} \leq \mu(A_S) \leq 1$. The lower bound is attained when $S_2 = S$ ($\pi_0 p_0 = \pi_1 p_1$ almost everywhere on S). In that case, the scaled densities are equal a.e. and the probability of being in class 0 or 1 is $\frac{1}{2}$, which means that on average, half of the observations are well classified.

[1] This definition is a variant of the original definition of the total variation [BGvdM92, GS02]

Corollary 1
The score $\mu(A_S)$ is bounded by the total variation as follows:

$$\begin{cases} \max\left(\frac{1}{2}, \frac{1-|\pi_0-\pi_1|}{2} + (1+|\pi_0-\pi_1|)\frac{\delta_{TV}(p_0,p_1)}{2}\right) \leq \mu(A_S) \\ \mu(A_s) \leq \min\left(1, \frac{1+|\pi_0-\pi_1|}{2} + \frac{\delta_{TV}(p_0,p_1)}{2}\right) \end{cases}.$$

The proof is given in Section 5.5 in the Appendix.

Remark 4
Many properties arise from Corollary 1:

1. *If $p_0 = p_1$ almost everywhere, we have: $\frac{1}{2} \leq \mu(A_S) \leq \frac{1+|\pi_0-\pi_1|}{2}$.*
2. *If $\pi_0 = \pi_1$, we have: $\mu(A_S) = \frac{1}{2} + \frac{\delta_{TV}(p_0,p_1)}{2}$.*
3. *If $\pi_i = 1$ (then $\pi_j = 0$ for $j \neq i$), we have:*
   $\max(\frac{1}{2}, \delta_{TV}(p_0,p_1)) \leq \mu(A_S) \leq 1$.
4. *If $\delta_{TV}(p_0,p_1) = 1$, we have: $\mu(A_S) = 1$.*

The result of the Proposition presented above states that the success rate of the classifier using the reduced input x can be directly related to the total variation of the projected densities. Aiming at quantifying the loss with respect to the classifier that exploits at best all the dictionary entries, we prove the following result:

Proposition 2
Let ρ_0 and ρ_1 be the densities defined in Equation (6.1). Then, it holds:

$$\mu(A_S) \leq \min\left(1, \frac{1+|\pi_0-\pi_1|}{2} + \frac{\delta_{TV}(\rho_0,\rho_1)}{2}\right).$$

The proof of Proposition 2 in given in Section 5.5 in the Appendix. The result of the above Proposition shows that the total variation in the dictionary space of dimension n_g is a natural upper bound of the score. By projecting on a Stiefel manifold we cannot improve with respect to the best classifier that uses all the information.

An illustration of the relationships between the score and the total variation is proposed in Figure 5.1.

From the inequality shown on $\mu(A_S)$ in Corollary 1, many relations can be established with other metrics [GS02]. In this paper we compare the success rate measure to the Hellinger distance (see 5.2.1.2) and the symmetrised Kullback-Leibler divergence (see 5.2.1.3).

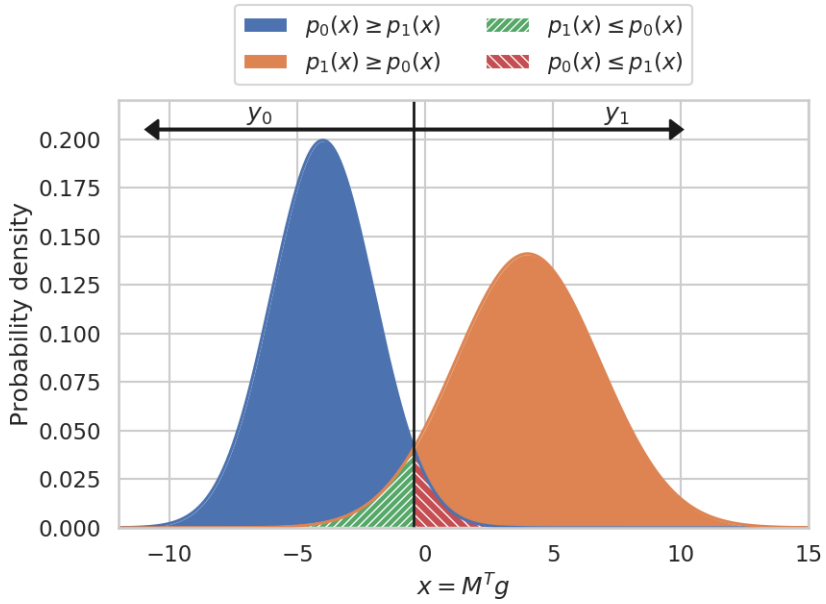

Figure 5.1: Example of the pdf of two classes (0 and 1) in the projected space. Here, $\pi_0 = \pi_1 = \frac{1}{2}$ and $\mu_L(S_2) = 0$ (the Lebesgue measure of S_2).

5.2.1.2 Relation to the Hellinger distance

The Hellinger distance (see Definition 10) is a f-divergence [Csi64]. For some studies the Hellinger distance is preferred to other common f-divergences as the Kullback-Leibler divergence or χ^2-divergence which are not metrics [She14].

Definition 10
The Hellinger distance d_H between two absolutely continuous probability distributions P and Q on S, with pdfs p and q respectively is:

$$d_H^2(P,Q) \triangleq \frac{1}{2}\int_S \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx.$$

Using the following inequalities between the Hellinger distance and the total variation (see [Sur21] for the proofs):

$$d_H^2(P_0,P_1) \leq \delta_{TV}(p_0,p_1) \leq \sqrt{2}d_H(P_0,P_1),$$

we can establish the following result:

Proposition 3
Let P_0 and P_1 be two probability distributions on S and p_0 and p_1 the corresponding pdf. Then,

$$\begin{cases} \max\left(\frac{1}{2}, \frac{1-|\pi_0-\pi_1|}{2} + (1+|\pi_0-\pi_1|)\frac{d_H^2(P_0,P_1)}{2}\right) \leq \mu(A_S) \\ \mu(A_S) \leq \min\left(1, \frac{1+|\pi_0-\pi_1|}{2} + \frac{\sqrt{2}d_H^2(p_0,p_1)}{2}\right) \end{cases},$$

where d_H is the Hellinger distance (see Definition 10) and $\mu(A_S)$ the success event measure defined previously in Equation (5.5).

The proof of Proposition 3 is immediate using the result of the Corollary 1 and the inequalities between the Hellinger distance and the total variation, proposed in [DP17].

5.2.1.3 Relation to the symmetrised Kullback-Leibler divergence

The Kullback-Leibler divergence (or relative entropy) [KL51] (see Definition 11) is a measure of the dissimilarity of a probability distribution to another. It reads:

Definition 11
The Kullback-Leibler divergence between two continuous probability distributions P and Q on S, with pdf p and q respectively is:

$$D_{KL}(P||Q) \triangleq \int_S p(x) \ln\left(\frac{p(x)}{q(x)}\right) dx.$$

In many classification problems, for symmetry reasons, the symmetrised Kullback-Leibler divergence is introduced: $D_{SKL}(P,Q) \triangleq \frac{1}{2}\big(D_{KL}(P||Q) + D_{KL}(Q||P)\big)$. Aiming at improving the classification, the maximisation of the symmetrised Kullback-Leibler divergence is proposed [Big03, LS03, LFGS16, RSB+04]. Hereafter, a result is proved relating the classification score defined in Equation (5.5) to the symmetrised Kullback-Leibler divergence.

Proposition 4
Let P_0 and P_1 be two continuous probability distributions on S (see Definition of the set S in 7) with pdf p_0 and p_1 respectively.

If $\log\left(\frac{p_0}{p_1}\right) \in L^\infty(S)$ and, moreover, $D_{KL}(p_i||p_j) < +\inf$ for $i \neq j, i,j = 0$ or 1 (absolute continuity of P_i with rescpect to P_j) then, $\exists c > 0$ such that the following inequalities hold:

$$2\Big(2\mu(A_S) - (1+|\pi_0-\pi_1|)\Big)^2 \leq D_{SKL}(P_0,P_1) \leq c\frac{2\mu(A_S) + |\pi_0-\pi_1| - 1}{1+|\pi_0-\pi_1|}.$$

Under the hypothesis of Propostion 4, we clearly see that the minimum of $D_{SKL}(P_0,P_1)$ is 0 and it is reached for $\mu(A_S) = \frac{1}{2}$ (the minimum of $\mu(A_S)$). Moreover, increasing the success rate is equivalent to increase the value of the symmetrised Kullback-Leibler divergence.

5.2.1.4 Some words on the semi-supervised classification

It may appear that labelled data belong to only one class whereas more than one class exist in the classification task. In this context, a classical supervised classification will fail as all the samples will have the same label as the one considered at the learning phase. To overcome this, we first introduce a dummy class for which the pdf is described in Equation (5.7)

$$p_{dummy} = \pi_0 p_0 + \pi_1 p_1. \tag{5.7}$$

This dummy class corresponds to either the labelled samples or unlabelled samples. We now consider the mixture between the pdf of labelled samples (let say 0) and the pdf of either labelled and unlabelled samples as shown in Equation (5.8).

$$p'(x) = \pi'_0 p'_0(x) + \pi'_1 p'_{dummy}(x). \tag{5.8}$$

Let $\mu'(A_S)$ be the corresponding new score. Then from Equation (5.8) the pdf of the dummy class described in Equation (5.7) and the definition of the score in Equation (5.5) we have:

$$\mu'(A_S) = \frac{1}{2} + \frac{1}{2}\int_S |\pi'_0 p_0 - \pi'_1(\pi_0 p_0 + \pi_1 p_1)| dx.$$

As we have less information in semi-supervised classification, it is reasonable to set the *a priori* π_i and π'_i to $\frac{1}{2}$. It immediately yields to:

$$\mu'(A_S) = \mu(A_S) - \frac{1}{4}\delta_{TV}(p_0,p_1). \tag{5.9}$$

Under these assumptions, the result in Equation (5.9) is expected. With less information (semi-supervised case) we cannot expect to have a better score than in the supervised scenario. Moreover, the score on the semi-supervised case directly depends on the score on the supervised case and the total variation between the pdf of the two classes.

5.2.2 Optimisation of the classification success rate

The method proposed consists in choosing an element of the Stiefel manifold to define the input of the classifier: $x = M^T g$. The goal is to optimise the score $\mu(A_S)$ introduced and commented in the section above. Optimising over all the possible elements of the Stiefel manifolds (of multiple and unknown dimension k) would be prohibitive. To circumvent this, a double greedy approach is proposed. A comprehensive analysis of the possible formulation of greedy methods and their analysis is proposed in [Tem15].

The heuristics we follow are the following: the smaller the dimension of the input, the better it is in terms of palliating the curse of dimensionality; aiming at reducing possible overfitting phenomena, the sparser the orthonormal vectors of M, the better it is. Henceforth, the strategy which is investigated is the following: we start with $k = 1$ and look for a vector of unitary norm such that at each step of a greedy method, we maximise

$\mu(A_S)$. When the error on a validation set stagnates and start increasing (early stopping criterion [Pre98]), we start considering $k = 2$. The first column vector of M is the result of the previous step of the method, and by a greedy approach we construct a second unitary norm column vector, orthogonal to the first one. This can be iterated until the error on a validation set starts increasing as soon as we start building the $(k+1)-$th vector.

5.2.2.1 Computation of $\mu(A_S)$

Before detailing the double greedy algorithm in Section 5.2.2.2, let us introduce a strategy to approximate the measure of the success events $\mu(A_S)$. In general, the densities p_0 and p_1 are not known. Instead, samples are given. To approximate the integral in Equation (5.5), we use a Montecarlo approach: in the present case, it turns out to be a counting of how many samples are correctly classified, that is $y = y_*$). The difficulty is to precisely estimate the regions S_0, S_1 and S_2. For that, an estimation of the values of p_0, p_1 is required. Since the dimension k is usually small (for instance $k = 1,2,3$), a Kernel Density Estimation (KDE) is a viable way to estimate the values of p_0 and p_1 and hence to have an approximation of the decomposition of S. For larger values of k, KDE could become impractical and costly from a numerical point of view [LW19]. A surrogate is proposed, based on the use of the Mahalanobis distance [DMJRM00, XNZ08]. This provides a perfect outcome in the case of Gaussian distributions. Since, in general, the projected densities p_0 and p_1 are not Gaussians, an approximation based on hierarchical clustering is proposed. Roughly speaking, classes i ($i = 0,1$) may be seen as a mixture of Gaussian distributions of means $(\mu_i^{(1)},\dots,\mu_i^{(l)})$ and covariance matrices $(\Sigma_i^{(1)},\dots,\Sigma_i^{(l)})$, that can be computed by clustering. For an observation x with a label $y_* = i$ the success event s is given by:

$$d_i^{(k)} = (x-\mu_i^{(k)})^T[\Sigma_i^{(k)}]^{-1}(x-\mu_i^{(k)}),\ i = 0,1,$$
$$s(x) = \begin{cases} 1 \text{ if } \min_{k=1,\dots,l_i} d_i^{(k)} < \min_{k=1,\dots,l_j} d_j^{(k)} \text{ and } y_* = i\ (j \neq i) \\ 0 \text{ otherwise} \end{cases}. \tag{5.10}$$

For all the entries of the dataset, the individual score s proposed in Equation (5.10) can be evaluated. The approximation of $\mu(A_S)$ to be used reads:

$$\mu(A_S) \approx \frac{\pi_0}{n_0}\sum_{l=1}^{n_0} s(x^{(l)}_{|y_*=0}) + \frac{\pi_1}{n_1}\sum_{l=1}^{n_1} s(x^{(l)}_{|y_*=1}),$$

where $n_0 + n_1 = n_s$, with n_0 and n_1 are the number of samples labelled $y_* = 0$ and $y_* = 1$ respectively. This is an empirical approximation of the score introduced in Equation (5.5). The error introduced by such an approximation and possible alternatives are discussed in [BCDD14]. An example of score estimation is shown in Figure 5.2. In this example, the distribution of the class 0 is a mixture of a Gaussian and a uniform

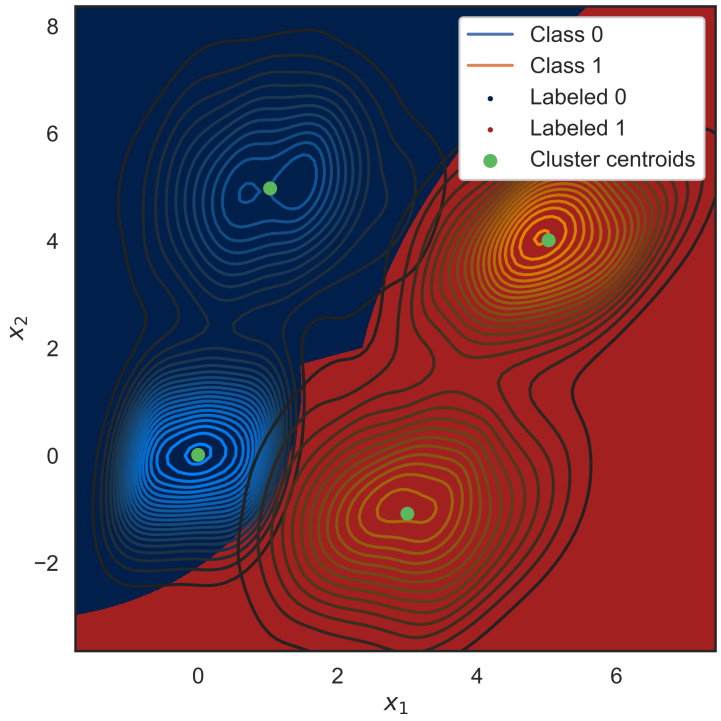

Figure 5.2: Example of classification using Mahalanobis distance. The Kernel Density Estimation using a Gaussian kernel shows the distribution for the two classes. Cluster centroids were obtained using DBSCAN. Class 0: uniform distribution on the square centred on 0 and a side of length 1 and a bivariate Gaussian distribution with $\mu = (1,5)$ and identity covariance matrix. Class 1: Gaussian bivariate distributions $\mu_a = (5,4)$ and $\mu_b = (3,-1)$ with $\Sigma_a = (0.8, 0.2; 0.2, 0.6)$ and $\Sigma_b = Id$. Sample size of 500 for each distribution.

distribution; class 1 is a mixture of two Gaussian distributions. Samples are drawn and the hierarchical clustering algorithm applied.

The bound of the probability of being in class i is then given by the multivariate Chebyshev inequality [Nav13].

5.2.2.2 DGDR algorithm

Let n_s,$n_v \in \mathbb{N}^*$ be the number of the samples used in the training and the validation phases respectively. A training and a validation datasets $(g^{(i)},y_*^{(i)})_{i=1}^{n_s}$, $(g^{(i)},y_*^{(i)})_{i=1}^{n_v}$ are given, that consist of couples of dictionary entries and corresponding labels.

Let $\widehat{M}_{k,n_g} \in \mathcal{M}_{k,n_g}$ be the element of the Stiefel Manifold selected at the $k-$th outer iteration of the method. The goal is to find a vector $\omega_* \in \mathbb{R}^{n_g}$, orthogonal to all the columns of the matrix $\widehat{M}_{k,n_g}$, such that:

$$\widehat{M}_{k+1,n_g} = [\widehat{M}_{k,n_g}, \omega_*],$$
$$x \in \mathbb{R}^{k+1},\ x = \widehat{M}^T_{k+1,n_g} g,$$
$$\omega_* = \arg \sup_{\omega \in \mathbb{R}^{n_g}} \mu(A_S).$$

When n_g is large, this optimisation can be costly. Furthermore, when the vector ω is sparse the classification tends to be less prone to overfitting phenomena. For these reasons, ω is constructed in a greedy way. At first $||\omega||_{\ell^{0,n_g}} = 1$, so that only one dictionary entry is chosen, by computing the value of the score (on the training dataset) for all possible choices and keeping the best.

At the beginning of the $l-$th inner iteration, $||\omega||_{\ell^{0,n_g}} = l-1$, $l-1$ dictionary entries have been chosen and we have to choose the $l-$th one. Let the chosen indices be in the set $c^{(k+1)} = \{i_1, \ldots, i_{l-1}\}$. The $l-$th non-zero entry has to be chosen among the indices $i \in c_c^{(k+1)}$, the complementary set of $c^{(k+1)}$. Moreover, the best values of the selected entries of ω are sought, such that the result of the classification is the best possible (in the sense of the score introduced). Once one candidate to be the $l-$th non-zero component is proposed, an optimisation task on the entries of ω is performed by using the CMAES method, detailed in [IB09, KMH+04]. This does not guarantee automatically that ω is orthogonal to the subspace spanned by the column of $\widehat{M}_{k,n_g}$. Otherwise stated, $[\widehat{M}_{k,n_g},\omega] \in Gr_{k+1,n_g}$. The projection onto the Stiefel manifold is obtained by QR decomposition. Let $Q_m \in \mathbb{R}^{n_g \times k+1}$, $R_m \in \mathbb{R}^{k+1 \times k+1}$, it holds:

$$Q_m R_m = [\widehat{M}_{k,n_g},\omega],$$
$$\widehat{M}_{k+1,n_g} = Q_m.$$

Among all the possible optimised choices for the $l-$th component, the one that maximises the score is chosen. As said, the stopping criterion for these iterations is the early stopping strategy [Pre98]: the score is computed on the validation set. A stagnation of the score ends the inner iteration. As soon as increasing the dimension of the Stiefel manifold does not produce an improvement on the score computed on the validation, the outer iterations end. Once the algorithm terminates, the element of the Stiefel manifold is obtained. Details of the method are shown in Algorithm 5.2.2.2.

Remark 5

When, in the proposed method, $||\omega||_{\ell^{0,n_g}} = 1$, $\forall k$, the Feature Selection (FS) [GE03] reduction is retrieved, as a particular case. Furthermore, when the objective function is not the quantity $\mu(A_S)$ but the ℓ^2 norm of the samples g reconstruction, the proposed approach turns out to be a sparse approximation of the Principal Component Analysis (PCA) of the data (a description is provided in [Bis06, WEG87]). The outcome of the proposed method is therefore a set of orthonormal modes that does not coincide with the PCA modes. These two methods, FS and PCA, are the most used dimension reduction

techniques when dealing with classification problems. A numerical comparison will be proposed in Section 5.3.

Algorithm 1 DGDR algorithm

$k \leftarrow 1$; $c \leftarrow [1,\ldots,n_g]$ {Dimensional counter; Dictionary entry indices.}
$\mu(A_S)_v^{new} \leftarrow 1/2$ {Minimal reachable value of $\mu(A_S)$ for the validation set.}
$\mu(A_S)_v^{old} \leftarrow 0$ {Success event measure of the validation set†.}
$\widehat{M} \leftarrow [\,]$ {Empty matrix which will be an element of $\mathbf{Gr}_{k,n_g}$.}
while $\mu(A_S)_v^{new} > \mu(A_S)_v^{old}$ **do**
  $j \leftarrow 1$; $c^{(k)} \leftarrow [\,]$ {Component counter; Stores selected entry indices.}
  while $\mu(A_S)_v^{new} > \mu(A_S)_v^{old}$ **do**
    $\mu(A_S)_v^{old} \leftarrow \mu(A_S)_v^{new}$ {Update stop criteria.}
    $\mu \leftarrow [0]^{n_g}$;$W \leftarrow [0]^{n_g \times n_g}$ {To store scores;weights.}
    $c_c^{(k)} \leftarrow c \backslash c^{(k)}$
    for $l \in c_c^{(k)}$ **do**
      Initialize ω_l {Initialize non-zeros indices for CMAES.}
      $\mu(A_S)_l$,$\omega_l \leftarrow CMAES(\omega_l, (g_i,y_i^*)_{i=1}^{n_s})$ {ω_l optimisation††.}
      $\mu_l \leftarrow \mu(A_S)_l$ {Assign the l^{th} component of μ.}
      $\omega \leftarrow Weights(\omega_l, s_j, l)$ {Generate l^{th} weight column vector of $W^{\ddagger}$.}
      $W_l \leftarrow \omega$ {Assign the l^{th} column of the weight matrix W.}
    end for
    $l_* \leftarrow \arg\max_l \mu_l$ {New dictionary entry position for the contribution.}
    $\omega_* \leftarrow W_{l_*}$ {Extract corresponding weights.}
    $\widehat{M}_* \leftarrow [\widehat{M},\omega_*]$; $M_* \leftarrow QR(\widehat{M}_*)$
    $\mathcal{D}_v \leftarrow (M_*^T g_i, y_i^*)_{i=1}^{n_v}$; $\mathcal{D}_t \leftarrow (M_*^T g_i, y_i^*)_{i=1}^{n_s}$ {Projected sets.}
    $\mu(A_S)_v^{new} \leftarrow Score(\mathcal{D}_v, \mathcal{D}_t)$ {Compute score on the validation set‡‡.}
    $s_j \leftarrow [s_j,l_*]$; $j \leftarrow j+1$
  end while
  $\widehat{M} \leftarrow [\widehat{M},\omega_*]$; $k \leftarrow k+1$
end while
return $\widehat{M}$

†: Any value lower than $\mu(A_S)_v^{new}$ to enter in the while loop.
††: For each ω_l computed at each CMAES step, the QR decomposition of $[\widehat{M},\omega_l]$ and projection of the training set $(x_i = M^T g_i, y_i^*)_{i=1}^{n_s}$ are performed to compute $\mu(A_S)_l$.
‡:$[0]^{n_g}$ vector with optimised weights assigned to the non-zero positions s_j and l.
‡‡: The validation score is computed using KDE or Mahalanobis distance through the projected training set.

5.2.3 Principle of analysis

In this section an analysis of the proposed method is presented. The goal is to show that, in the limit case of an infinite number of samples, or, in alternative, the perfect knowledge of the pdf, the proposed method tends to maximise the score. In the case in which all the dictionary entries are used, the score by exploiting all the entries is retrieved.

Proposition 5
Let $\widehat{M}_{k,n_g} \in \mathcal{M}_{k,n_g}$ and $\widehat{M}_{k,n_g} = [\widehat{M}_{k-1,n_g}, \omega]$; let $1 \leq m < n_g$, and $||\omega||_{\ell^{0,n_g}} = m$. The set of non-zero entries of ω is denoted by c, whose cardinality is $\#c = m$. Let $\tilde{\omega} \in \mathbb{R}^{n_g}$. The set of non-zero entries of $\tilde{\omega}$ is $\tilde{c}$, $\#\tilde{c} = m + 1$. It holds $c \subset \tilde{c}$. Then,

$$\max_{\tilde{\omega}} \mu(A_S) \geq \max_{\omega} \mu(A_S).$$

This proposition shows that, in the inner iteration, as far as we add terms to the vector ω, the score improves. The proof is immediate. The function $\mu(A_S)$ is bounded, the Stiefel manifold is a compact set, and the set of non-zero entries of ω is strictly included in the one of $\tilde{\omega}$. Henceforth, the conclusion. Indeed, at worst, the non-zero entry of $\tilde{\omega}$ which is not a non-zero entry of ω can be set to zero and the equality would hold. The outer iteration, the one in which the dimension of the element of the Stiefel manifold is increased, is the object of the following Proposition.

Proposition 6
Let $M_{k,n_g} \in \mathcal{M}_{k,n_g}$ and the associated score be $\mu(A_S^{(k)})$. Let $M_{k+1,n_g} \in \mathcal{M}_{k+1,n_g}$ such that:

$$M_{k+1,n_g} = [M_{k,n_g}, \omega],$$

where $\omega \in \mathbb{R}^{n_g}$ and the associated score be $\mu(A_S^{(k+1)})$. Then:

$$\mu(A_S^{(k+1)}) \geq \mu(A_S^{(k)}).$$

The demonstration of Proposition 6 is given in Section 5.5 in the Appendix.

Remark 6
Since, at each step of the method, we enforce that the matrices M_{k,n_g} belong to the Stiefel manifold, when $k = n_g$ we retrieve an element of the orthogonal group, whose associated score is the maximal possible.

5.3 Computational studies

In this section, we compare the algorithm with classical tools used for dimension reduction in the context of classification problems. The first part consists of comparing the strategy proposed in this paper with Feature Selection (FS) [GE03]. In the second part we make the comparison with the Principal Component Analysis (PCA) method [Bis06, WEG87].

5.3.1 Comparison with feature selection

FS is a widely used dimension reduction tool consisting of selecting a subset of features, pertinent to answer a clustering [DL00] or classification [DL97] problem. In the context of the present work, this would consist in selecting a subset of the dictionary entries, and, as remarked, it can be seen as a particular case of the proposed method.

In this first test case a synthetic example is constructed by considering a Gaussian mixture: $\rho_0(g)$,$\rho_1(g)$ are two normal distributions, of mean and variance (μ_0,Σ_0) and (μ_1,Σ_1) respectively. When dealing with Gaussian distributions, the symmetrised Kullback-Leibler divergence can be analytically computed. In Section 5.2.1, an equivalence between the symmetrised KL divergence and the score $\mu(A_S)$ is shown. The symmetrised Kullback-Leibler divergence between the distributions reads:

$$D_{SKL}(\rho_0,\rho_1) = \frac{1}{4}\Big(tr(\Sigma_1^{-1}\Sigma_0 + \Sigma_0^{-1}\Sigma_1) + (\mu_1-\mu_0)^T(\Sigma_0^{-1}+\Sigma_1^{-1})(\mu_1-\mu_0) - 2n_g\Big).$$

Let $k \in \mathbb{N}^*$ denote the number of entries selected by the FS, let $l \in \mathbb{N}^*, l \leq n_g$ be such that the elements of the Stiefel manifold are $(M_{l,n_g} \in \mathcal{M}_{l,n_g})$ and $m \in \mathbb{N}^*$ be the maximal number of non-zero entries of the columns of M_{l,n_g}.

When projecting Gaussian distributions on linear subspaces, Gaussian distributions are retrieved, namely P_0,P_1, whose densities are $p_0(x)$,$p_1(x)$. The mean and variances of these are reported in Table 5.3.1 for the case of FS and the proposed method.

Classification strategy	Σ_0	Σ_1	μ_0	μ_1
FS	I_k	$\beta I_k, \beta > 0$	0_k	1_k
DGDR	I_l	$\beta I_l, \beta > 0$	0_l	$(\sqrt{m},\ldots,\sqrt{m}) \in \mathbb{R}^l_+$

Table 5.1: Section 5.3.1: Gaussian parameters for feature selection and double greedy algorithm study case.

The symmetrised Kullback-Leibler divergence for FS and, respectively, for the double greedy dimension reduction algorithm (DGDR) reads:

$$D^{(FS)}_{SKL}(P_0,P_1) = \frac{1}{4}\left(\frac{k}{\beta} + k\beta + k(1+\frac{1}{\beta}) - 2k\right),$$

$$D^{(DGDR)}_{SKL}(P_0,P_1) = \frac{1}{4}\left(\frac{l}{\beta} + l\beta + lm(1+\frac{1}{\beta}) - 2l\right).$$

An analysis of the above expressions provides some insight on the performances of the methods. Let us consider the difference between the divergences:

$$f_{l,k}(\beta) = \left(\frac{k}{l} - 1\right)\beta^2 + \left(2 - \frac{k}{l}\right)\beta + 2\frac{k}{l} - 1, \tag{5.11}$$

$$\Delta = D^{(DGDR)}_{SKL}(P_0,P_1) - D^{(FS)}_{SKL}(P_0,P_1) \geq 0 \Longleftrightarrow m \geq \frac{f_{l,k}(\beta)}{\beta+1}. \tag{5.12}$$

Some properties are highlighted:

- If $k = l$, then, $\forall\beta$, the symmetrised KL divergence is larger for DGDR if $m \geq 1$; in the case in which $m = 1$, as commented before, the methods coincide.
- If $k < l$, $\forall\beta$, $\Delta \geq 0$: in this case DGDR always outperforms FS.
- if $k > l$, different scenarios are possible.
- It is interesting to consider the case of identical Gaussians, namely $\beta = 1$, the DGDR outperforms FS if $m > \frac{k}{l}$. Remark that when $l = 1$ (DGDR selects just an element of the unit sphere): $\Delta \geq 0$ if $m \geq k$.

In general, when both the methods achieve the same result in terms of symmetrised KL divergence, DGDR method has a reduced dimension smaller (in some cases much smaller) than FS. This is particularly relevant when a finite (and not so large) number of samples are available. A comparison is given in Table 5.3.1 where the symmetrised Kullback-Leibler difference between DGDR and FS is computed for some values of k, l and m.

Δ		$m=1$					$m=5$					$m=10$				
		l					l					l				
		1	5	10	15	20	1	5	10	15	20	1	5	10	15	20
k	1	0.0	2.0	4.5	7.0	9.5	2.0	12.0	24.5	37.0	49.5	4.5	24.5	49.5	74.5	99.5
	5	-2.0	0.0	2.5	5.0	7.5	0.0	10.0	22.5	35.0	47.5	2.5	22.5	47.5	72.5	97.5
	10	-4.5	-2.5	0.0	2.5	5.0	-2.5	7.5	20.0	32.5	45.0	0.0	20.0	45.0	70.0	95.0
	15	-7.0	-5.0	-2.5	0.0	2.5	-5.0	5.0	17.5	30.0	42.5	-2.5	17.5	42.5	67.5	92.5
	20	-9.5	-7.5	-5.0	-2.5	0.0	-7.5	2.5	15.0	27.5	40.0	-5.0	15.0	40.0	65.0	90.0

Table 5.2: Section 5.3.1: Δ difference between symmetrised Kullback-Leibler divergences obtained using DGDR (parameters l and m) and FS (parameter k) defined in Equation (5.11). Here, the covariance matrix factor β is set to 1 (see Table 5.3.1).

5.3.2 Comparison with PCA

In this section, we compare the proposed double greedy algorithm with the Principal Component Analysis (PCA), which consists in finding an orthogonal transformation such that the variance of the dataset in the principal directions is the largest possible [Bis06, WEG87].

Let G follows a multivariate normal distribution in $\mathbb{R}^{n_g}$ defined by its covariance matrix $\Sigma = \mathbf{I}_{n_g}$ and its mean $\mu = \mathbf{0}_{n_g}$. Let $l \in \mathbb{N}^*$ and $I = \{i_1, \ldots, i_l\}$ be a set of indices. For this test case, classes 0,1 are defined as:

$$\begin{cases} y^* = 0 \text{ if } g_{i_1}, \ldots, g_{i_l} \geq 0, \\ y^* = 1 \text{ otherwise.} \end{cases}$$

In the particular case analysed hereafter, $n_g = 50$ and $I = \{10,11,12,13\}$. The total number of samples in the training set is $n_s = 1500$, 105 for the class 0 and 1395 for the

class 1. The number of samples for the validation set is $n_v = 500$, 35 samples for the class 0 and 465 for the class 1. The DGDR method led to $m = 2$ dictionary entries to build the first direction and $m = 3$ dictionary entries to build the second direction.

In comparison, the PCA method was applied on the same training set as for the DGDR method (i.e. same samples starting from the same dimension $n_g = 50$) in such a way that the output dimension is the same as the one obtained with the DGDR method (i.e. 2 directions).

In Figure 5.3, the samples projected on the first two principal directions obtained by PCA are shown. As it can be assessed, PCA does not provide an efficient pre-processing, the conditional densities of the classes being practically indistinguishable.

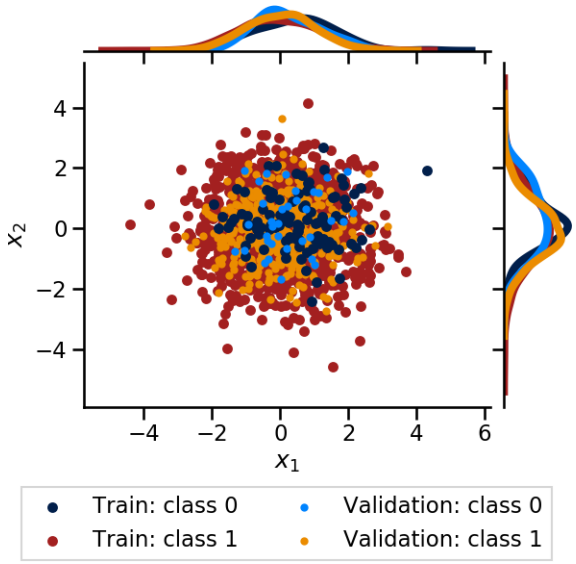

Figure 5.3: Samples projected on the first two directions computed by PCA, along with the marginal conditional densities.

On the other hand, the first two directions identified by the Double Greedy method proposed (M_{2,n_g}) tend to maximise the separation between the conditional densities. The samples projected on these directions are shown in Figure 5.4.

The results obtained allow to stress an important aspect. PCA is a general purpose reduction method, which is often effective, but it is not specific to classification tasks, as the method proposed. Henceforth, there are situations, like the one shown in this example, in which PCA fails in providing a well performing dimension reduction.

5.3.3 Comparison with metric learning techniques

For this study, the same classification task and dataset as described in Section 5.3.2 are given. The double greedy algorithm (DGDR) is compared with Averaged Neighbourhood Margin Maximisation (ANMM), Neighbourhood Component Analysis (NCA) and

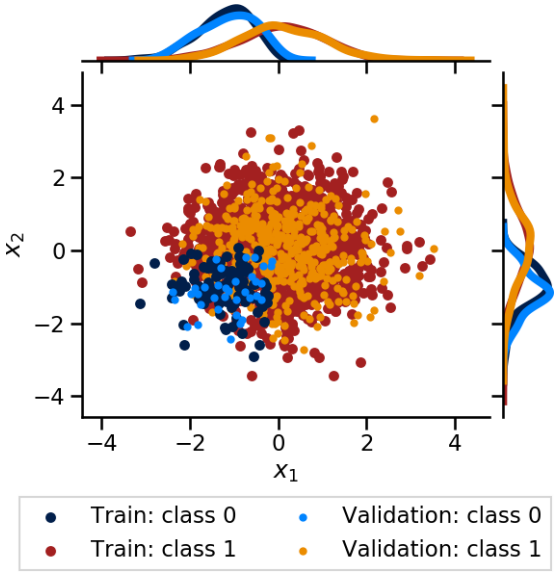

Figure 5.4: Samples projected on M_{2,n_g} obtained by the double greedy approach and the associated marginal conditional densities. The Mahalanobis distance was used for the classification (see Section 5.2.2.1). Using the early-stopping criterion on the validation set, two components were chosen for the first dimension and three for the second.

Partial Least Squares (PLS) dimension reduction techniques. For the comparison, all the dimension reduction tools are set such that the reduced space is bi-dimensional (see Figure 5.5).

For the same subspace dimension, the DGDR technique shows a better separation between the two classes. The projection in the first direction is approximately the same for the DGDR, ANMM and PLS techniques. However, the projection in the second direction results in a better discrimination of the densities (of each class) for the DGDR technique.

5.3.4 A high-dimensional low sample size example

We consider a numerical illustration of a high-dimensional low sample size regime. For this $n_g = 10^5$, and the number of samples in the training set is $n_s = 200$, evenly distributed between the two classes. The validation set consists of $n_v = 100$ samples. Let $I = 10, \ldots, 20$ be a set of indices, whose cardinality is $\#I = 11$. The probability density function reads:

$$\rho(g) = \frac{1}{2}\rho_0(g) + \frac{1}{2}\rho_1(g),$$

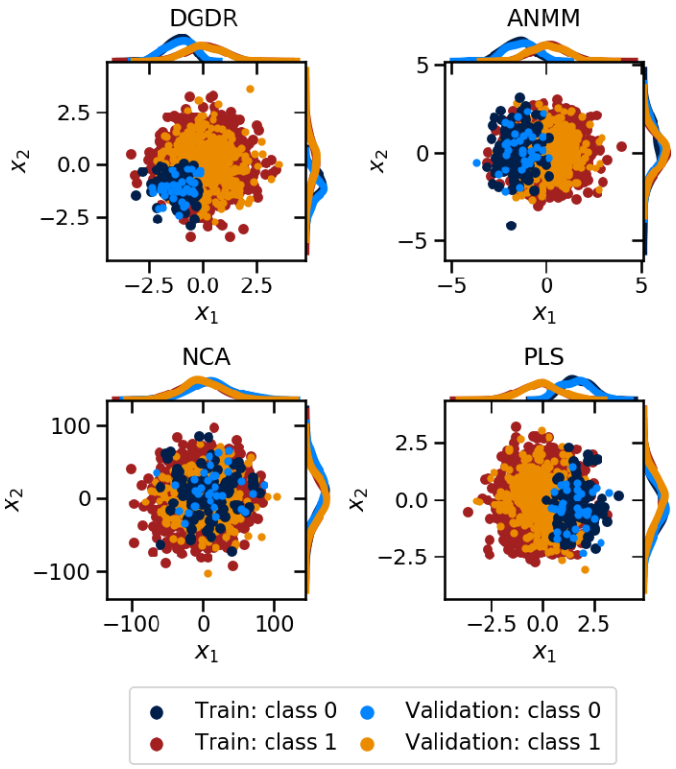

Figure 5.5: Samples projected onto a bi-dimensional subspace for different dimension reduction techniques: Double Greedy method (DGDR), Averaged Neighbourhood Margin Maximisation (ANMM), Neighbourhood Component Analysis (NCA) and Partial Least Squares (PLS).

where ρ_0 and ρ_1 are unitary variance Gaussians, whose mean are $\mu_0 = [0, \ldots, 0]$ and:

$$\begin{cases} \mu_{1i} = \eta \text{ if } i \in I, \\ \mu_{1i} = 0 \text{ otherwise.} \end{cases}$$

For this example, we considered two cases, namely $\eta = 1$ and $\eta = 4$; the Mahalanobis distance criterion was used to approximate the score.

The results are shown in Figure 5.3.4, when the dimension of the reduced space is $k = 1$. The samples are projected in $x \in \mathbb{R}$ and the probability densities of the two classes are plotted for the training and validation sets. In the upper row, $||\omega||_{\ell^{0,n_g}} = 1$, in the lower row $||\omega||_{\ell^{0,n_g}} = 2$. Visually, we can assess that the separation between the densities increases when we use two components instead of one, and this holds for both the training and the validation sets; this is confirmed by the increase in the classification score. When $\eta = 4$, the densities of the two classes ρ_0, ρ_1 have a larger total variation with respect to the case $\eta = 1$. This is also found for the marginal densities p_0,p_1. The two non-zero components chosen by the algorithm to construct the first direction (M_{1,n_g}) are elements

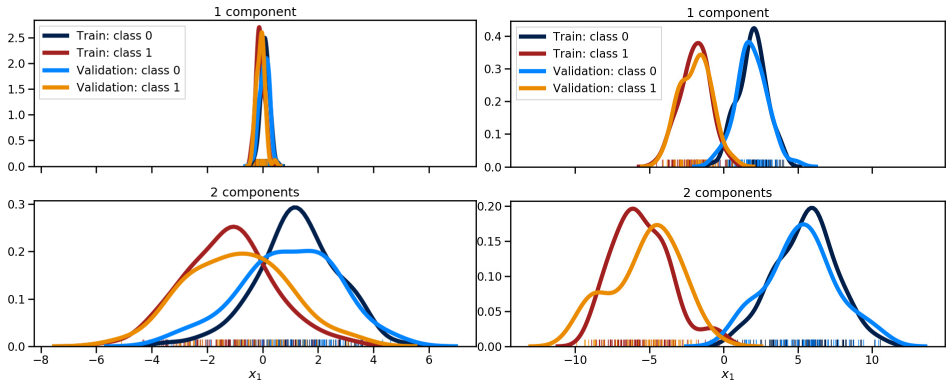

Figure 5.6: Test case in Section 5.3.4, projected distributions on $x \in \mathbb{R}$ for the case $\eta = 1$ (left panel) and $\eta = 4$ (right panel). Upper row, $||\omega||_{\ell^{0,n_g}} = 1$, lower, $||\omega||_{\ell^{0,n_g}} = 2$.

of I described above. Selected non-zero components are 12 and 10 for $\eta = 1$ study case and 14 and 11 for $\eta = 4$.

5.3.5 Application to classification problems

To conclude, we present two tests based on realistic datasets. For these studies, the data used were treated without any pre-processing stage. The classification was performed with four different classifiers available in the Scikit-Learn library [PVG+11], by using parameters by default. These classifiers are: Discriminant Analysis (LDA), K-Nearest Neighbours (KNN), Decision Tree (DT), Naive Bayes (NB) and Support Vector Machines (SVM).

5.3.5.1 LSVT voice rehabilitation

The LSVT Voice Rehabilitation data set is provided by UCI machine learning repository[2]. This dataset based on dysphonia measures is studied to assess the LSVT protocol in Parkinson's disease [TLFR13], and was used, as other datasets in this repository, as a benchmark to test several classification strategies.

It consists of a sample of size $n_s = 126$, for a features dictionary of $n_g = 309$ entries. The output is the voice rehabilitation, labelled "Acceptable" or "Unacceptable". For this binary classification, we trained on the first $n_t = 100$ samples and validate on the last $n_v = 26$ samples.

The proposed DGDR method stopped providing an input dimension $k = 2$. The number of components selected to produce by linear combination x_1,x_2 are $m_1 = 2$ and $m_2 = 2$. All the dimension reduction tools (DRT) are set such that the reduced space is bi-dimensional.

The success rates on the validation set for the two dimension reduction strategies and the five classifiers are given in Table 5.3.5.1.

[2] https://archive.ics.uci.edu/ml/index.php

		Classifier					
		LDA	KNN	DT	NB	SVM	GPC
DRT	DGDR	0.846	0.846	0.846	0.846	0.923	0.846
	ANMM	0.692	0.538	0.615	0.346	0.692	0.308
	NCA	0.692	0.538	0.538	0.308	0.692	0.308
	PCA	0.692	0.538	0.731	0.385	0.692	0.308
	PLS	0.692	0.538	0.692	0.385	0.692	0.308

Table 5.3: Section 5.3.5.1: Classification success rates for the same input dimension (2).

Regardless of the classifier, the DGDR technique always gives a better classification success rate. ANMM and NCA returned a better success rate than DGDR with LDA but for a subspace of dimension 27 (0.846) and 15 (0.962) respectively.

5.3.5.2 Wisconsin breast cancer

The Wisconsin breast cancer dataset is also provided by UCI machine learning repository[3]. It consists of $n_s = 569$ samples and a dictionary of $n_g = 31$ entries. The output is the breast cancer diagnosis, labelled "Malignant" or "Benign". For this binary classification, we trained on the first $n_t = 400$ samples and validated on the last $n_v = 169$ samples.

Using the early stopping criterion, the DGDR technique stopped at an input dimension $k = 3$. The number of components used to construct the inputs x_1, x_2 and x_3 are respectively $m_1 = 17$, $m_2 = 14$ and $m_3 = 1$. The comparison is the same as in the previous section (see Section 5.3.5.1). Classification success rates on the validation set are given in Table 5.3.5.2.

		Classifier					
		LDA	KNN	DT	NB	SVM	GPC
DRT	DGDR	0.976	0.935	0.923	0.929	0.941	0.941
	ANMM	0.935	0.769	0.899	0.781	0.769	0.769
	NCA	0.864	0.763	0.852	0.769	0.769	0.769
	PCA	0.935	0.769	0.876	0.781	0.769	0.769
	PLS	0.935	0.769	0.893	0.781	0.769	0.769

Table 5.4: Section 5.3.5.2: Classification success rates for the same input dimension (3).

For this study, the dimension reduction using the DGDR technique results in a higher classification success rate than using the others dimension reduction techniques. Only LDA and DT classifiers have returned a higher classification success rate, but for larger values of the subspace dimension (see Table 5.3.5.2). In particular, these values were always $k \leq 4$.

[3] https://archive.ics.uci.edu/ml/index.php

		Classifier	
		LDA	DT
DRT	ANMM	0.982(**14**)	0.923(**4**)
	NCA	0.988(**23**)	0.959(**12**)
	PCA	0.988(**20**)	0.935(**6**)
	PLS	0.988(**15**)	lower

Table 5.5: Section 5.3.5.2: Best classification success rates with the corresponding subspace dimension **(X)**.

5.4 Conclusion

This paper investigates a double greedy algorithm to construct the input x of a classifier by exploiting a large number of dictionary entries. The method is designed to deal with classification problems in a high-dimensional/low sample size regime. The method can be interpreted as a sparse goal oriented dimension reduction technique. The first contribution is the introduction of an objective function to be maximised, which is directly related to the performances of the classifiers in the reduced space. This objective function was related to quantities which are commonly used to assess the performances in classification problems. The method proposed is easily parallelisable and hence well adapted to large problems. Some examples are proposed to illustrate the performances of the proposed method: first, a comparison in a small-scale problem is performed with Feature Selection and the Principal Component Analysis; then, the method was tested on a large scale synthetic example that mimics a high-dimensional/low sample size regime and a realistic dataset.

Several perspectives arise. One concerns the application of the method to a broader set of realistic cases. The extension to more than two classes as well as to regression problems will be considered.

5.5 Appendix

Proof of Proposition 1.

A first relationship between the score and the densities is derived from the normalisation condition. As $p(x)$ is a density, it holds:

$$\int_S p(x)dx = 1$$

$$\implies \int_{S_0} \pi_0 p_0 + \pi_1 p_1 dx + \int_{S_1} \pi_0 p_0 + \pi_1 p_1 dx + \int_{S_2} \pi_0 p_0 + \pi_1 p_1 dx = 1,$$

by the properties of the measures and the definition of $p(x)$. The terms of the definition of $\mu(A_S)$ are isolated, providing:

$$\mu(A_S) + \frac{1}{2}\int_{S_2} p(x)dx + \int_{S_0} \pi_1 p_1(x)dx + \int_{S_1} \pi_0 p_0(x)dx = 1. \tag{5.13}$$

Second, by adding and subtracting the same terms to the definition of the score, aiming at highlighting its relationship with the total variation we have:

$$\mu(A_S) = \int_{S_0} (\pi_0 p_0 - \pi_1 p_1)dx + \int_{S_1} (\pi_1 p_1 - \pi_0 p_0)dx + \int_{S_0} \pi_1 p_1 dx + \frac{1}{2}\int_{S_2} p dx.$$

By making the use of the result in Equation (5.13), we get:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big(\int_{S_0} (\pi_0 p_0 - \pi_1 p_1)dx + \int_{S_1} (\pi_1 p_1 - \pi_0 p_0)dx\Big).$$

It holds that on S_0 we have $\pi_0 p_0 - \pi_1 p_1 > 0$ and the converse holds on S_1, almost everywhere. Moreover, on S_2 it holds that $\pi_0 p_0 - \pi_1 p_1 = 0$. Henceforth:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big(\int_S |\pi_0 p_0 - \pi_1 p_1|dx\Big).$$

Which concludes the proof.

■

Proof of Corollary 1.

We start with the second inequality. From the definition of S_0 and S_1, we have:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big(\int_{S_0} |\pi_0 p_0 - \pi_1 p_1|dx + \int_{S_1} |\pi_1 p_1 - \pi_0 p_0|dx\Big).$$

We immediately have $\mu(A_S) \leq 1$. Let $\epsilon = \frac{\pi_0 - \pi_1}{2}$. Then,

$$|\pi_0 p_0 - \pi_1 p_1| = |(\frac{1}{2} + \epsilon)p_0 - (\frac{1}{2} - \epsilon)p_1| = |\frac{1}{2}(p_0 - p_1) + \epsilon(p_0 + p_1)|.$$

By the use of the triangular inequality, we obtain:

$$|\pi_0 p_0 - \pi_1 p_1| \leq \frac{1}{2}|p_0 - p_1| + |\epsilon|(p_0 + p_1).$$

Then,

$$\mu(A_S) \leq \frac{1}{2} + \frac{1}{2}\Big(\frac{1}{2}\int_{S_0 \cup S_1} |p_0 - p_1|dx + |\epsilon| \int_{S_0 \cup S_1} (p_0 + p_1)dx\Big).$$

Recalling that the first term in the parenthesis is lower than the total variation between p_0 and p_1 and the second term in the parenthesis is majored by $2|\epsilon|$ as p_0 and p_1 are probability density function, for the second inequality we have:

$$\mu(A_S) \leq \min\Big(1, \frac{1 + |\pi_0 - \pi_1|}{2} + \frac{\delta_{TV}(p_0,p_1)}{2}\Big).$$

For the left part of the inequality, by definition of the sets, it is clear that $\mu(A_S) > \frac{1}{2}$. We assume that $\pi_0 \geq \pi_1$ and pose $\epsilon = \frac{\pi_0 - \pi_1}{2}$. By symmetry, we can proceed in the same way for $\pi_1 \geq \pi_0$. Then,

$$\mu = \frac{1}{2} + \frac{1}{2}\Big[\int_{S_0} \Big((\frac{1}{2} + \epsilon)p_0 - (\frac{1}{2} - \epsilon)p_1\Big)dx + \int_{S_1} \Big((\frac{1}{2} - \epsilon)p_1 - (\frac{1}{2} + \epsilon)p_0\Big)dx\Big].$$

$$\Longleftrightarrow$$

$$\mu = \frac{1}{2} + \frac{1}{2}\Big[\frac{1}{2}\int_{S_0}(p_0 - p_1)dx + \frac{1}{2}\int_{S_1}(p_1 - p_0)dx + \epsilon\int_{S_0}(p_0 + p_1)dx - \epsilon\int_{S_1}(p_0 + p_1)dx\Big]. \quad (5.14)$$

We recall the definition of S_0 and S_1:

$$\begin{cases} S_0 = \{x \in \mathbb{R}^{n_g} | \pi_0 p_0(x) > \pi_1 p_1(x)\} \\ S_1 = \{x \in \mathbb{R}^{n_g} | \pi_1 p_1(x) > \pi_0 p_0(x)\} \end{cases}.$$

Let S_0' and S_1' the two following sets:

$$\begin{cases} S_0' = \{x \in \mathbb{R}^{n_g} | p_0(x) > p_1(x)\} \\ S_1' = \{x \in \mathbb{R}^{n_g} | p_1(x) > p_0(x)\} \end{cases}.$$

As $\pi_0 > \pi_1$ and p_0 and p_1 are non-negative (since they are probability density functions), we have the following inclusions:

$$\begin{cases} S_0' \subseteq S_0 \\ S_1 \subseteq S_1' \end{cases}.$$

Let:

$$\begin{cases} \widehat{S}_0 = \{x|(\pi_0 p_0 > \pi_1 p_1) \wedge (p_0 = p_1)\} \\ \widetilde{S}_0 = \{x|(\pi_0 p_0 > \pi_1 p_1) \wedge (p_0 < p_1)\} \end{cases} \text{and} \begin{cases} \widehat{S}_1 = \{x|(\pi_0 p_0 = \pi_1 p_1) \wedge (p_0 < p_1)\} \\ \widetilde{S}_1 = \{x|(\pi_0 p_0 > \pi_1 p_1) \wedge (p_0 < p_1)\} \end{cases}.$$

Then, $S_i = S_i' \cup \widetilde{S}_i \cup \widehat{S}_i$ for $i \in \{0,1\}$, is the union of disjoint sets. We denote $\widetilde{S} = \widetilde{S}_0 = \widetilde{S}_1$. Thanks to the disjointedness of the sets, we can rewrite each term of $\mu(A_S)$ in Equation (5.14) as follows:

$$\begin{cases} \frac{1}{2}\int_{S_0}(p_0 - p_1)dx = \frac{1}{2}\Big[\int_{S_0'}(p_0 - p_1)dx + \int_{\widehat{S}_0}(p_0 - p_1)dx + \int_{\widetilde{S}}(p_0 - p_1)dx\Big] \\ \epsilon\int_{S_0}(p_0 + p_1)dx = \epsilon\Big[\int_{S_0'}(p_0 + p_1)dx + \int_{\widehat{S}_0}(p_0 + p_1)dx + \int_{\widetilde{S}}(p_0 + p_1)dx\Big] \\ \frac{1}{2}\int_{S_1}(p_1 - p_0)dx = \frac{1}{2}\Big[\int_{S_1'}(p_1 - p_0)dx - \int_{\widehat{S}_1}(p_1 - p_0)dx - \int_{\widetilde{S}}(p_1 - p_0)dx\Big] \\ -\epsilon\int_{S_1}(p_0 + p_1)dx = \epsilon\Big[-\int_{S_1'}(p_0 + p_1)dx + \int_{\widehat{S}_1}(p_0 + p_1)dx + \int_{\widetilde{S}}(p_0 + p_1)dx\Big] \end{cases}.$$

Then,

$$\begin{cases} \mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big\{\frac{1}{2}\Big[\int_{S_0'}(p_0 - p_1)dx + \int_{S_1'}(p_1 - p_0)dx\Big] + \epsilon\int_{S_0'}(p_0 + p_1)dx \\ -\epsilon\int_{S_1'}(p_0 + p_1)dx + \int_{\widetilde{S}}\Big(2\epsilon(p_0 + p_1) - p_1 + p_0\Big)dx \\ +\int_{\widehat{S}_0}\Big(\frac{p_0 - p_1}{2} + \epsilon(p_0 + p_1)\Big)dx + \int_{\widehat{S}_1}\Big(\epsilon(p_0 + p_1) - \frac{p_1 - p_0}{2}\Big)dx\Big\}. \end{cases}$$

The first term in the brackets is exactly the total variation between p_0 and p_1. This is because $p_0 - p_1 > 0$ on S_0', $p_1 - p_0 > 0$ on S_1', $p_0 = p_1$ on S_2' and $S = \cup_{i=0}^{2} S_i'$ and they are disjoint. For term four in the bracket, we have:

$$2\epsilon(p_0 + p_1) - p_1 + p_0 = 2(\pi_0 p_0 - \pi_1 p_1),$$

which is positive on $\widetilde{S}$. For term five in the bracket, we know that $p_0 = p_1$ on $\widehat{S}_0$. Then, the first term in this integral is equal to 0. Finally, for the last term in the bracket, we have:

$$\epsilon(p_0 + p_1) - \frac{p_1 - p_0}{2} = \pi_0 p_0 - \pi_1 p_1.$$

However, as this term is integrated on $\widehat{S}_1$, this term is equal to 0. It follows that:

$$\begin{aligned} \mu(A_S) = \frac{1}{2} + \frac{1}{2}\Big\{\delta_{TV}(p_0,p_1) &+ \epsilon\int_{S_0'}(p_0 + p_1)dx - \epsilon\int_{S_1'}(p_0 + p_1)dx \\ &+ \epsilon\int_{\widehat{S}_0}(p_0 + p_1)dx + 2\int_{\widetilde{S}}(\pi_0 p_0 - \pi_1 p_1)dx\Big\}. \end{aligned}$$

We have:

$$\begin{cases} \epsilon\int_{S_0'}(p_0 + p_1)dx = \epsilon\int_{S_0'}(p_0 - p_1)dx + 2\epsilon\int_{S_0'} p_1 dx \\ -\epsilon\int_{S_1'}(p_0 + p_1)dx = \epsilon\int_{S_1'}(p_1 - p_0)dx - 2\epsilon\int_{S_1'} p_1 dx \end{cases}.$$

It follows that:

$$\epsilon \int_{S'_0} (p_0 + p_1)dx - \epsilon \int_{S'_1} (p_0 + p_1)dx = 2\epsilon\delta_{TV}(p_0,p_1) + 2\epsilon \int_{S'_0} p_1 dx - 2\epsilon \int_{S'_1} p_1 dx.$$

Then,

$$\mu(A_S) \geq \frac{1}{2} + (1 + 2\epsilon)\frac{\delta_{TV}(p_0,p_1)}{2} - \epsilon$$

Finally,

$$\begin{cases} \max\left(\frac{1}{2}, \frac{1-|\pi_0-\pi_1|}{2} + (1 + |\pi_0 - \pi_1|)\frac{\delta_{TV}(p_0,p_1)}{2}\right) \leq \mu(A_S) \\ \mu(A_S) \leq \min\left(1, \frac{1+|\pi_0-\pi_1|}{2} + \frac{\delta_{TV}(p_0,p_1)}{2}\right) \end{cases}.$$

Which concludes the proof.

■

Proof of Proposition 2.

Let M be an element of the Stiefel manifold $\mathcal{M}_{k,n_g}$, and $M^\perp$ its orthogonal complement of M. The score $\mu(A_S)$, by exploiting the change of coordinates and the properties of the elements of the orthogonal group, can be rewritten as follows:

$$\mu(A_S) = \frac{1}{2} + \frac{1}{2}\left(\int_{M_k} \left| \int_{M_k^\perp} (\pi_0\rho_0 - \pi_1\rho_1) d\xi_{k+1} \dots d\xi_{n_g} \right| d\xi_1 \dots d\xi_k \right).$$

By triangular inequality, we can write:

$$\mu(A_S) \leq \frac{1}{2} + \frac{1}{2}\left(\int_{M_k} \int_{M_k^\perp} |\pi_0\rho_0 - \pi_1\rho_1| \, d\xi_{k+1} \dots d\xi_{n_g} d\xi_1 \dots d\xi_k \right).$$

$$\mu(A_S) \leq \min\left(1, \frac{1 + |\pi_0 - \pi_1|}{2} + \frac{\delta_{TV}(\rho_0,\rho_1)}{2}\right).$$

■

Proof of Proposition 4.

The right-hand side inequality is proved by making use of the Pinsker inequality [FHT03]:

$$\delta^2_{TV}(P_0,P_1) \leq \frac{1}{2} D_{SKL}(P_0,P_1).$$

Then, using the inequality between the total variation distance and the measure of the success events in Corollary 1, we directly get:

$$D_{SKL}(P_0,P_1) \geq 2\Big(2\mu(A_S) - (1 + |\pi_0 - \pi_1|)\Big)^2.$$

To prove the inequality on the left-hand side we consider the definition of the symmetrised Kullback-Leibler divergence:

$$D_{SKL}(P_0,P_1) = \frac{1}{2}\int_S (p_0 - p_1) \ln\left(\frac{p_0}{p_1}\right) dx.$$

As, $\log\left(\frac{p_0}{p_1}\right) \in L^\infty(S)$, Hölder inequality leads to:

$$D_{SKL}(P_0,P_1) \leq \frac{||\log(p_0/p_1)||_{L^\infty}}{2} \int_S |p_0 - p_1| dx.$$

In what follows we set: $c = ||\log(p_0/p_1)||_{L^\infty}$. The definition of the total variation is inserted:

$$D_{SKL}(P_0,P_1) \leq c\delta_{TV}(P_0,P_1).$$

Then,

$$D_{SKL}(P_0,P_1) \leq c\frac{2\mu(A_S) + |\pi_0 - \pi_1| - 1}{1 + |\pi_0 - \pi_1|},$$

which concludes the proof. ∎

Proof of Proposition 6.

Let us denote $h = \pi_0\rho_0 - \pi_1\rho_1$, S_k the space obtained by projecting $g \in \mathbb{R}^{n_g}$ onto the columns of M_{k,n_g}. Let its orthogonal complement be denoted by $S_k^\perp$. The element of the orthogonal group constructed from M_{k,n_g} is denoted by $R = [M_{k,n_g}, M_{k,n_g}^\perp]$. It holds:

$$\xi = R^T g,$$
$$x = [\xi_1; \ldots; \xi_k].$$

As remarked in Equation (5.2), $p(x)$ is obtained by:

$$p(x) = \int_{S_k^\perp} \rho(\xi) d\xi_1, \ldots d\xi_k.$$

The score (and the total variation) is then directly related to the following integral:

$$I^{(k)} = \int_{S_k} \left| \int_{S_k^\perp} h \; d\xi_{k+1} \ldots d\xi_{n_g} \right| d\xi_1 \ldots d\xi_k.$$

Without loss of generality let us suppose that:

$$\xi_{k+1} = \omega^T g.$$

Remark that the orthogonal complement to S_k can be always constructed in this way. We will denote by $S_{k+1}^\perp = S_k^\perp / \xi_{k+1}$. Hence:

$$I^{(k)} = \int_{S_k} \left| \int_{-\infty}^{\infty} \left(\int_{S_{k+1}^\perp} h \; d\xi_{k+2} \ldots d\xi_{n_g} \right) d\xi_{k+1} \right| d\xi_1 \ldots d\xi_k.$$

When ω is used to construct the input ($x = M_{k+1,n_g}$), the integral $I^{(k+1)}$ reads:

$$I^{(k+1)} = \int_{S_k} \int_{-\infty}^{\infty} \left| \int_{S_{k+1}^\perp} h \; d\xi_{k+2} \ldots d\xi_{n_g} \right| d\xi_{k+1} d\xi_1 \ldots d\xi_k.$$

A straightforward inequality follows:

$$\int_{-\infty}^{\infty} \left| \int_{S_{k+1}^{\perp}} h \; d\xi_{k+2} \dots d\xi_{n_g} \right| \; d\xi_{k+1} - \left| \int_{-\infty}^{\infty} \left(\int_{S_{k+1}^{\perp}} h \; d\xi_{k+2} \dots d\xi_{n_g} \right) d\xi_{k+1} \right| \geq 0,$$

which implies:

$$\mu(A_S^{(k+1)}) - \mu(A_S^{(k)}) = \frac{1}{2} \left(I^{(k+1)} - I^{(k)} \right) \geq 0,$$

and this concludes the proof. ■

Chapter 6

A method to enrich experimental datasets by means of numerical simulations in view of classification tasks

Classification tasks are frequent in many applications in science and engineering. A wide variety of statistical learning methods exists to deal with these problems. However, in many industrial applications, the number of available samples to train and construct a classifier is scarce and this has an impact on the classifications performances. In this work, we consider the case in which some *a priori* information on the system is available in form of a mathematical model. In particular, a set of numerical simulations of the system can be integrated to the experimental dataset. The main question we address is how to integrate them systematically in order to improve the classification performances. The method proposed is based on Nearest Neighbours and on the notion of Hausdorff distance between sets. Some theoretical results and several numerical studies are proposed.

Contents

6.1 Introduction

Classification tasks are frequent in many applications in science and engineering. The statistical learning methods which are proposed to deal with them rely on the fact that many examples (where the number of samples depends on the application under consideration) are available and can be exploited to uncover the underlying structure of the data and their separation in several classes. After the learning phase has been performed, a classifier is set up and can be used to infer to which class a new observed sample belongs to.

In many industrial applications the number of available samples is scarce, impacting the performances of the classification. A way to circumvent this limitation is to integrate to the available *a posteriori* information (provided by the available data) some *a priori* information (coming from experimental insight or theoretical knowledge) as proposed for instance in [MFH+18, MHB96, Joh03, HKC+04].

The use of mathematical models and numerical simulations to construct the training set of machine learning methods has been recently investigated in [TPYP18, BH20, RMMC20]. In [TPYP18], a model order reduction framework is proposed in order to deal with classification problems. In this, synthetic outputs obtained by numerical simulations are used in order to train the machine learning algorithms. The influence of the model error on the classification performance is investigated. In [BH20], numerical simulations are used to set up a sparse gaussian process. This is used in order to solve an optimal design problem for structural anomaly detection. In [RMMC20], a convolutional neural network framework is proposed to efficiently deal with health monitoring, seen as a classification problem on multivariate time series. The training of the network is performed by using numerical simulations of a physical based model of the system.

In this work we consider the case in which some *a priori* information is available in form of a mathematical model. Numerical simulations of several instances of the model can be computed and integrated to an available dataset in order to improve the classification performances. The main questions to be answered are: how many numerical simulations should we include, and which ones? Which information is needed in order to devise a systematic strategy? This work is devoted to the investigation of possible answers to these questions, in the spirit of what has been proposed in [BM10], in which an adaptive sampling is proposed in order to improve the performances of a SVM classifier. The selection of the samples aims at improving the position of the support vectors and the margin. These questions have also been raised in [LHS14], where each training sample is weighted in order to solve SVM classification tasks.

This topic is also closely related to two research fields in machine learning: domain adaptation and instance (or prototype) selection. The main goal of domain adaptation is to account for the discrepancies between target and test sets and propose ways to correct for them. An abundant literature on this subject is available, [WD18, SSW15, PGLC15, ZSMW13]. The main difference with respect to the method proposed in the present work consists in the fact that in domain adaptation we often try to minimise a discrepancy between the datasets, whereas in the present work we focus on trying to

improve a classification score. This is more similar, in the spirit, to the methods proposed in the field of instance selection. Different kinds of algorithms have been proposed in this research field and can be divided into 4 different classes (commented and compared in the recent work [BK20]):

1. Incremental, such as Condensed Nearest Neighbours [Har68] and its variants [RWLI75, VSP05] or Instance-based learning [AKA91]. These methods consist in building the training set by adding samples, chosen according different criteria.

2. Decremental such as Decremental Reduction Optimisation Procedure [WM97, WM00] or Hit Miss Network [Mar08] consist in defining the training set by pruning samples from an available reservoir of potentially redundant (and corrupted) samples.

3. Batching such as Edited Nearest Neighbours [T+76], consists in testing whether each sample of the training set follows a removable criterion. All of the samples verifying this criterion are removed at once.

4. Fixed size such as Learning Vector Quantisation [NE14] which consists in fixing *a priori* the size of the training set and selecting the samples to be used.

Recent studies have proposed in-between methods such as in [CHL05]. These algorithms might have several drawbacks: in the methods in which we test one sample at a time and we decide if it has to be included or not into the training set, we might obtain a result which is sensitive to the order with which we test the samples. In some methods, the fitness function introduced to perform the selection is based on similarity criteria applied to the input features rather than the classification success rate, which might be suboptimal in some cases or it might depend upon hyperparameters which need to be tuned.

The main contributions of the present investigation are the following:

1. A systematic strategy can be set up, that enrich available training sets and improves the classification performance in a substantial way. The only information which is exploited is a representative validation set, given even in form of samples or in form of a set of data and parameters of a reliable mathematical model describing the phenomenon.

2. The method which is proposed can be decomposed in two phases: an incremental one, in which we add to the training set samples taken from a reservoir of numerical simulations; a decremental one in which we prune samples to reduce redundancy and noise oversensitivity. We tried to reduce as much as possible the number of hyperparameters.

3. The obtained approach is not a generative one: it is not strictly needed to have an exhaustive training set distributed as the validation set; it is sufficient to add the most informative samples, in a sense that will be made more precise in the

following, and that will be encoded in the fit functions used in the incremental and decremental phases.

The structure of the work is as follows. In Section 6.2 the method is proposed, and some properties are investigated from a theoretical standpoint. In Section 6.3 the discretisation is discussed, and in Section 6.4 some numerical test cases are presented to illustrate the approach.

6.2 Method

In this section, we detail the method proposed in the present work. The problem under investigation is a classification task, and, for the sake of simplicity, we restrict to a binary classification. Four different sets of samples are introduced:

1. An *augmented set*, for which we know both the input (observations) and the output (labels). The augmented set is the main unknown of the problem. We wish to devise a way to construct it, starting from an available scarce (in the number of samples) set of labelled instances. The training set of the problem (we will use to set up the classifier) is the augmented set at the end of the enrichment process. The elements of the augmented set will be denoted by the superscript "tr".

2. A *validation* set, for which we know both the input (observations) and the output (labels), whose elements will be denoted by the superscript "v". This is the only source of information to construct the augmented set.

3. A *test* set, for which we know just the observation, whose elements will be denoted by the superscript "te".

4. A *reservoir* of numerical simulations of the systems, for which we know the observation and the label, to be used in order to construct or enrich the augmented set.

Several possible cases are met in realistic applications. First, we can be in a case in which we have an available experimental dataset covering all the possible meaningful instances of the problem under scrutiny, having however, not so many samples (or not enough to have the wished performance on the test set). We will call this a *complete validation* case. Second, we could be in an *incomplete validation* case, meaning that the experimental dataset to be used as training and validation cover only a subset of the possible instances (occurring in the test set). In both these situations, we would like to enrich the dataset by integrating elements of the reservoir in the augmented set. This is the simplest way to integrate some *a priori* information coming from mathematical modelling to the existing *a posteriori* information of the experimental data. We will consider here the cases of a perfect model (useful to validate certain aspect of the method) and the more realistic case in which the model is biased.

6.2.1 Context and notations

Let X be a random variable, representing the state of a system, for a population of individuals. A system configuration, identified by the realisation x, can belong to two classes, labelled $y = \{0,1\}$. In an application, the system is observed through a measurement process and for a given observation $g \in \mathbb{R}^{n_g}$ (which in general results from the application of a non-linear function to x), we need to uncover whether the state belongs to the class $y = 0$ or $y = 1$.

The system observable for the population can be modelled by a random variable X_{n_g} defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$, with $\Omega \subseteq \mathbb{R}^{n_g}$, $\mathcal{A}$ the σ-algebra of all the possible observable and $\mathbb{P}$ the probability measure. We denote $g \in \Omega$ a realisation of X_{n_g} and we assume that its probability density distribution, denoted $\rho(g)$, is a mixture of two densities. Let π_0,$\pi_1 \in (0,1)$, such that $\pi_0 + \pi_1 = 1$. The probability density distribution reads:

$$\rho(g) = \pi_0\rho_0(g) + \pi_1\rho_1(g), \tag{6.1}$$

where $\rho_0(g)$ and $\rho_1(g)$ are the conditional probability density distributions for the classes 0 and 1 respectively, namely $\rho_{0,1}(g) = \rho\big(g|y = (0,1)\big)$.

Remark 7
Note that the above assumption is exactly the same as the one described in the DGDR method (see Section 5.1.1).

In the following, the Lebesgue measure of a generic set A is denoted by $\mu_L(A)$. The classification success rate is based on a score function μ_s, which is a measure, introduced and described in the DGDR method 5.2.1, and that we recall for sake of completeness. The set of all the subsets in Ω is denoted by 2^Ω.

Definition 12
We define the score function μ_s as follows:

$$\mu_s : \begin{cases} 2^\Omega \times 2^\Omega \ \rightarrow \ \mathbb{R}^+ \\ (S_0,S_1) \ \mapsto \mu_s(S_0,S_1) \end{cases},$$

where we take:

$$\mu_s(S_0,S_1) = \pi_0 \int_{S_0} \rho_0^s dg + \pi_1 \int_{S_1} \rho_1^s dg,$$

with the given densities ρ_0^s and ρ_1^s, and the superscript "s" denotes either the validation or the test set.

This score can be evaluated for all pairs of subsets S_0 and S_1. It is related to the classification outcome when we compute it for the following pair:

$$\begin{cases} S_0 = \{g \in \mathbb{R}^{n_g}, \pi_0\rho_0^{tr}(g) > \pi_1\rho_1^{tr}(g)\} \\ S_1 = \{g \in \mathbb{R}^{n_g}, \pi_1\rho_1^{tr}(g) > \pi_0\rho_0^{tr}(g)\} \end{cases},$$

where "tr" stands for the augmented set. As in the DGDR method (see Section 5.2.1), we make the following assumption:

$$\mu_L(S_2) = \mu_L\Big(\{g \in \mathbb{R}^{n_g}, \pi_1\rho_1^{tr}(g) = \pi_0\rho_0^{tr}(g)\}\Big) = 0.$$

Under the hypothesis that the set S_2 is a zero measure set, it follows that:

$$\rho_i^s = \rho_i^s \mathbf{1}_{S_i}, \forall i \Longrightarrow \mu_s = 1.$$

Remark 8
The main goal is to enrich the augmented set aiming at improving the classification performance, which is quantified by the above introduced score. To this end, it is not needed to have the following strong outcome:

$$\pi_i\rho_i^{tr} = \pi_i\rho_i^v,\ i \in \{0,1\}.$$

The propose approach is not a generative one seeking at generating samples distributed as the validation set, but samples which help improve the score. Henceforth, we could hopefully come up with a method which is less costly from a computational point of view.

6.2.2 Augmented set enrichment based on the Hausdorff distance: ASE-HD

We assume that Ω (defined in Section 6.2.1) is a measurable non-empty compact set of $\mathbb{R}^{n_g}$, and an observation of a system is $g \in \Omega \subset \mathbb{R}^{n_g}$.

At the beginning, the augmented set is given by the union of two known sets: $S_0^{(0)}$ and $S_1^{(0)}$: a sample of the augmented set is henceforth $g^{(tr)} \in S_0^{(0)} \cup S_1^{(0)}$. The goal is to progressively enrich the augmented set by making use of the samples in the reservoir of simulations. For the sake of simplicity, in this section, we make the hypothesis that the reservoir samples can cover Ω.

The information to be exploited comes from the knowledge of the validation set, either in form of samples or as a set of data and parameters of a mathematical model. This can be translated into two sets: $S_{0,1}^*$, with $S_1^* = \Omega \setminus S_0^*$, such that $S_0^* = \{g^{(v)} \in \Omega | y = 0\}$. These sets are optimal in the sense of the score function μ_v:

$$[S_0^*,S_1^*] = \arg \sup_{S_0,S_1 \subset \Omega} \mu_v.$$

In the following, we denote μ_* the score corresponding to these sets.

Let $n \in \mathbb{N}$ denotes the $n-$th step of the enrichment, we define $S_i^{(n)} \subseteq \Omega$ (for $i = 0$ or 1), the samples of the augmented set being $g^{(tr)} \in S_0^{(n)} \cup S_1^{(n)}$, as follows:

$$S_1^{(n)} = \Omega \setminus S_0^{(n)}. \tag{6.2}$$

The score of the classification corresponding to these sets reads:

Definition 13

$$\mu_v^{(n)} = \pi_0 \int_{S_0^{(n)}} \rho_0^v dg + \pi_1 \int_{S_1^{(n)}} \rho_1^v dg,$$

with:

$$\begin{cases} S_0^{(n)} = \{g \in \Omega, \pi_0 \rho_0^{(n)} > \pi_1 \rho_1^{(n)}\} \\ S_1^{(n)} = \{g \in \Omega, \pi_1 \rho_1^{(n)} > \pi_0 \rho_0^{(n)}\} \end{cases},$$

where $\rho_i^{(n)}$ is the pdf of the augmented set of class i and ρ_i^v is the pdf of the validation set of class i.

Starting from known sets $S_i^{(0)}$, $i = 0{,}1$, the goal is to transform them in order to converge to S_i^*, $i = 0{,}1$, which maximises the classification success rate. We construct a sequence which aims at increasing the cost function $\mu_v^{(n)}$, by observing that it is possible to make the sets $S_i^{(n)}$ converge towards the optimal sets S_i^* by diminishing a suitable distance between these sets.

Let $\mathcal{B}(g,\varepsilon) \subset \Omega$ denotes a ball of centre g and radius $\varepsilon \geq 0$. The enrichment method is performed as follows. Let $S_{0,1}^{(n)}$ be the available set estimations.

1. Define $M^{(n)} = (S_0^* \cap S_1^{(n)}) \cup (S_1^* \cap S_0^{(n)})$.
2. Solve the following problem[1]:

$$[g_{n+1}, \varepsilon_*] = \arg \sup_{g,\varepsilon \in \Omega} \left\{ \varepsilon \ | \mathcal{B}(g,\varepsilon) \subseteq M^{(n)} \right\}.$$

3. Let $\mathcal{B}_* = \mathcal{B}(g_{n+1}, \varepsilon)$. The update of the union of the intersections reads:

$$M^{(n+1)} = M^{(n)} \setminus \mathcal{B}_*,$$
$$S_0^{(n+1)} = \begin{cases} S_0^{(n)} \cup \mathcal{B}_* & \text{if } \mathcal{B}_* \subseteq S_0^* \cap S_1^{(n)} \\ S_0^{(n)} \setminus \mathcal{B}_* & \text{if } \mathcal{B}_* \subseteq S_1^* \cap S_0^{(n)} \end{cases}. \tag{6.3}$$

6.2.2.1 Analysis of the ASE-HD algorithm

The convergence of the sets $S_{0,1}^{(n)}$ to the sets $S_{0,1}^*$ is studied. First, a Lemma is introduced, clarifying the meaning of the set $M^{(n)}$. Let $A \Delta B$ be the symmetric difference [GH08] between the sets A and B.

Lemma 1
For the set $M^{(n)}$, $\forall n \in \mathbb{N}$ it holds:

$$M^{(n)} = S_0^* \Delta S_0^{(n)} = S_1^* \Delta S_1^{(n)}.$$

[1] On centrally symmetric sets, this would correspond to quantify the Bernstein widths of the set.

The result of this Lemma (demonstration given in Section 6.6 in the Appendix), makes it possible to prove the following result:

Proposition 7
Using the sequence of operations introduced above, almost surely, we have:

$$\lim_{n\to+\infty} \mu_v^{(n)} = \mu_*.$$

See the proof in Section 6.6 in the Appendix. Moreover, the gain on the score between two consecutive steps can easily be estimated. Its expression is given in the following result.

Corollary 2
Let $\mu_v^{(n)}$ be the score on the validation set at iteration $n \geq 0$. Then, $\forall n \in \mathbb{N}$, we have:

$$\mu_v^{(n+1)} - \mu_v^{(n)} = \int_{\mathcal{B}_*} |\pi_1 \rho_1^v - \pi_0 \rho_0^v| dg \geq 0,$$

with $\mathcal{B}_* = \mathcal{B}(g_{n+1}, \epsilon_*)$ defined in the previous section. Moreover, the equality holds if and only if $\mu_L(\mathcal{B}_*) = 0$, where μ_L denotes the Lebesgue measure.

See the demonstration in Section 6.6 in the Appendix. It follows that the gain is proportional to the total variation between ρ_0^v and ρ_1^v restricted to $\mathcal{B}_*$.

The result of the proposition states simply that, under the hypothesis that the system observable belongs to a compact set, and the set $S_{0,1}^*$ are known, the proposed iteration enriches the augmented set in such a way that the optimal classification score is retrieved. This algorithm shows some common properties with the algorithm detailed in [BCDD14]. In particular, the set sequence depends on the symmetric difference between the expected and the current set.

6.2.3 Reducing noise oversensitivity and bias induced errors: pruning

At each stage of the ASE-HD algorithm, the samples of the reservoir contained in a selected ball $\mathcal{B}_*$ are added to the augmented set (either to $S_0^{(n+1)}$ or to $S_1^{(n+1)}$). As remarked in [WM00], a large number of noisy samples could lead to noise oversensitivity. Moreover, as the augmented set is enriched through numerical simulations, a bias could potentially pollute the classification results in regions where the samples of the validation set are scarce. To avoid these phenomena and to make the classification less prone to overfitting, a pruning phase is introduced, which consists in removing the samples which are not useful in improving the score.

Once ASE-HD is performed, the obtained augmented set consists in the pair $S^{(n,0)} = (S_0^{(n)}, S_1^{(n)})$. Since, in practice, we have a finite number of samples, these sets consist in a finite set of balls centred around a finite number of samples.

A stochastic algorithm is introduced. At the j-th iteration, a sample $g_j \in S_0^{(n)} \cup S_1^{(n)}$ of the augmented set is randomly selected. It can be considered as the centre of a small

ball $\mathcal{B}_j(g_j,\varepsilon_j)$ whose radius ε_j is such that the other samples do not belong to $\mathcal{B}_j$. The score is computed and the following action is taken:

$$S^{(n,j+1)} = \begin{cases} S^{(n,j)} \setminus \mathcal{B}_j \text{ if } \mu_v(S^{(n,j)} \setminus \mathcal{B}_j) \geq \mu_v(S^{(n,j)}) \\ S^{(n,j)} \text{ otherwise} \end{cases} .$$

Remark that, by construction, at the end of the pruning step the score is at least as good as the beginning of the pruning step, and in some cases an improvement is obtained.

6.2.4 On realistic scenarios

In many applications different concerns may arise, such as the possible bias in the mathematical model (and then the database) [GA18, Ted06] and the incomplete validation case. We recall that in the present work we consider incomplete a validation set which does not cover the whole observable space Ω. In this section, a set of results are proposed to deal with these two cases.

6.2.4.1 Biased database

In general, the database obtained through a collection of experiments and/or simulations may have a bias. Let S_i^{te}, $(i = 0 \text{ or } 1)$ denote the test set which is supposed to cover Ω, *i.e.* $S_0^{te} \cup S_1^{te} = \Omega$:

$$\begin{cases} S_0^{(te)} = \{g \in \mathbb{R}^{n_g} | \pi_0 \rho_0^* > \pi_1 \rho_1^*\} \\ S_1^{(te)} = \{g \in \mathbb{R}^{n_g} | \pi_1 \rho_1^* > \pi_0 \rho_0^*\} \end{cases} . \tag{6.4}$$

The samples from these sets are samples drawn from the true underlying densities. The sets identified by using the densities of the model are:

$$\begin{cases} S_0^{(m)} = \{g \in \mathbb{R}^{n_g} | \pi_0 \rho_0^m > \pi_1 \rho_1^m\} \\ S_1^{(m)} = \{g \in \mathbb{R}^{n_g} | \pi_1 \rho_1^m > \pi_0 \rho_0^m\} \end{cases} . \tag{6.5}$$

The densities $\rho_{0,1}^m$ are in general different from the true ones. This is due to the model bias, which is such that the difference in the model state is propagated in the model observable g and hence in the density ρ^m. This, in turn, affects the sets $S_{0,1}^{(m)}$.

We recall that the sets satisfy:

$$\begin{cases} S_0^{te,m} \cup S_1^{te,m} = \Omega \\ S_0^{te,m} \cap S_1^{te,m} = \emptyset \end{cases} .$$

We define the biased sets as follows:

$$\begin{cases} b_0 = S_0^m \cap S_1^{te} \\ b_1 = S_1^m \cap S_0^{te} \end{cases} .$$

The bias sets $b_{0,1}$ are quantifying, in a sense which is pertinent for the binary classification, the effect of the model bias.

Lemma 2
Let the sets $S_{0,1}^{te,m}$ be defined as in Equation (6.4)-(6.5).The following equalities hold:

$$\begin{cases} S_0^m = (S_0^{te} \cup b_0) \setminus b_1 \\ S_1^m = (S_1^{te} \cup b_1) \setminus b_0 \end{cases}.$$

The result of the Lemma 2 (proof shown in Section 6.6 in the Appendix) makes it possible to prove the following result on the classification score of the test set:

Proposition 8
Let the hypothesis of Lemma 2 hold. Let

$$\mu_b = \mu_{te}(S_0^m, S_1^m) = \int_{S_0^m} \pi_0 \rho_0^{te} dg + \int_{S_1^m} \pi_1 \rho_1^{te} dg,$$

be the score of the classification of the test set when the augmented set is defined by the model. The maximal score is represented by:

$$\mu_* = \mu_{te}(S_0^{te}, S_1^{te}).$$

It holds:

$$0 \leq \mu_b \leq \mu_*,$$

and, moreover:

$$\begin{cases} \mu_b = \mu_* \iff \mu_L(b_i) = 0, \text{ for } i \in \{0,1\} \\ \mu_b = 0 \iff S_i^m = S_j^{te} \text{ and } \rho_j^{te} = \rho_j^{te} \mathbf{1}_{\{S_j^{te}\}}, \text{ for } i,j \in \{0,1\},\ i \neq j \end{cases}.$$

The demonstration is given in Section 6.6 in the Appendix.

Remark 9
In the case where $S_i^m = \emptyset$, we have $\mu_b = \int_\Omega \pi_j \rho_j^{te} dg$, $i \neq j$. It is straightforward to observe that in the case where there is no bias, we have the equality. In practice, we do not know S_j^{te}. It means that, if we only train with the model (database) we will compute the score over S_j^m.

6.2.4.2 The Validation set partially covers the set of possible outcomes

In several situations it is possible to assess whether the validation set covers all the possible scenarios that could occur in the test set (even prior of receiving the test set). This is possible in particular when there is an underlying parametrisation of the system at hand, namely when the scenarios of interest are associated with values of data and parameters that characterise the solution of the models describing the phenomenon. Here, we consider that the validation set partially covers Ω when the validation set does not have enough instances, in the sense that there are meaningful scenarios of the real system

which are not represented in the validation set. This would translate in the following: if we trained a classifier by using the validation set, it won't be able to classify well some query samples of the test set.

When the validation set partially covers Ω (incomplete validation set) we can show that the score on the test set (which is supposed to cover Ω) is lower than the score obtained with a validation set covering Ω (see Proposition 9).

Lemma 3
Let, $S_0^s \cup S_1^s = \Omega$ such that $S_0^s \cap S_1^s = \emptyset$ (for $s = te$ or v). Then,

$$S_1^{te} \setminus S_1^v = S_0^v \setminus S_0^{te}.$$

The demonstration is given in Section 6.6 in the Appendix.

Proposition 9
We denote $S_j^s = \{g | \pi_j \rho_j^s > \pi_k \rho_k^s\}$ $(k \neq j)$, where $s = te$ (test set) or v (validation set). We denote μ_{te}^c (resp. μ_{te}^p) the test set score obtained with a complete (resp. incomplete) validation set. By complete, we assume that the distribution of ρ_j^{te} and ρ_j^v are the same. Then,

$$\mu_{te}^p \leq \mu_{te}^c.$$

The demonstration is given in Section 6.6 in the Appendix. In this scenario, we cannot use generative adversarial networks (GANs) [GPAM+14] to enrich the augmented set in regions which are not covered by the validation set. This is due to the fact that the discriminator has no information on the region where there are no validation samples.

To enrich the augmented set, we propose first to enrich the validation set by adding to it samples extracted from the reservoir such that the enriched validation set covers all the possible meaningful scenarios.

If some information on the model bias is available (a statistics on the model bias), we proceed as follows. Let the bias in the observation be a random variable G_b, whose realisations are denoted by $g_b \in \mathbb{R}^{n_g}$. A sample of the reservoir is randomly picked in the region which is not covered by the validation set, whose observation is an element $g^{(r)} \in \mathbb{R}^{n_g}$. Then, a sample to be added to the validation set is:

$$g^{(v)} = g^{(r)} - g_b,$$

and the associated label is $y^{(v)} = y^{(r)}$.

6.3 Discretisation of the method

When the enrichment method proposed in the previous section has to be applied to realistic cases, we need to account for the fact that the only available quantity is a set of labelled samples, which can be divided into training and validation sets. The method needs to be discretised in order to be practically implemented. Several elements need to be detailed. The first one is the estimation of the score function. Its computation requires a density estimation.

6.3.1 Density estimation in high-dimension

To estimate the score by using a Monte Carlo method, we need to estimate a density in correspondence to a sample, namely the value $\rho(g) \in \mathbb{R}^+$. This task may be cumbersome due to the high-dimensionality of the space. Several methods of non-parametric density estimation are proposed in the literature [Bro07, Fry77, QX18]. For the present work we consider as a starting point the k–nearest neighbours (KNN) estimation. In the KNN method, a tree-based algorithm subdivides the samples set into overlapping balls, each containing a fix number of samples, say $k_{nn} \in \mathbb{N}^*$ on a total number of $N \in \mathbb{N}^*$ samples. The density is usually estimated by making the assumption that the density is roughly constant in a ball, leading to:

$$\rho(g^{(i)}) \approx \frac{k_{nn}/N}{\text{vol}(\mathcal{B}_i)},$$

where $\mathcal{B}_i = \mathcal{B}(g^{(i)}, \varepsilon_i)$ and $\text{vol}(\mathcal{B}_i)$ is its volume, computed according to the metric chosen to select the neighbours. We will denote the ℓ^p distance between two elements (g_1, g_2) as $||g_1 - g_2||_{\ell^{p,ng}}$.

Remark 10
Following [GSS19], if we want to classify a given sample g_* by using the Bayes rules, assuming $\mathbb{P}(y = 0) = \mathbb{P}(y = 1)$ and $N_0 = N_1 = N$, we will obtain the following result.

Let:

$$\begin{cases} g_0^{(tr)} = \arg\inf_{g \in S_0^{(tr)}} ||g_* - g||_{\ell^{p,ng}} \\ g_1^{(tr)} = \arg\inf_{g \in S_1^{(tr)}} ||g_* - g||_{\ell^{p,ng}} \end{cases}.$$

Furthermore, let ε_0 and ε_1 be the radius of the balls centred around $g_0^{(tr)}$ and $g_1^{(tr)}$ respectively. The *a posteriori* probability reads:

$$\mathbb{P}(y = 0|g_*) = \frac{\varepsilon_1^{n_g}}{\varepsilon_1^{n_g} + \varepsilon_0^{n_g}}.$$

This means that the classification outcome only depends on the distance between the closest points in each class in the augmented set and their respective k_{nn}^{th} nearest neighbour. Figure 6.1 shows an example in which, by making use of this approach we wrongly classify a validation point. As the computed radius is lower for class 1 the validation point is labelled 1 instead of 0.

The issue shown in Figure 6.1 is mainly due to the assumption that the density is constant in the ball. We propose of replacing it by an approximation based on Gaussian radial basis functions (RBFs). Let us introduce $\omega_i \in \mathbb{R}$, $i = 1, \ldots, k_{nn}$; moreover, let the elements in a ball be $g^{(i)} \in \mathbb{R}^{n_g}$, $i = 1, \ldots, k_{nn}$ and $\varepsilon_i > 0$ be the radius of the balls the samples $g^{(i)}$ are the centre of. The density in a ball is expressed as:

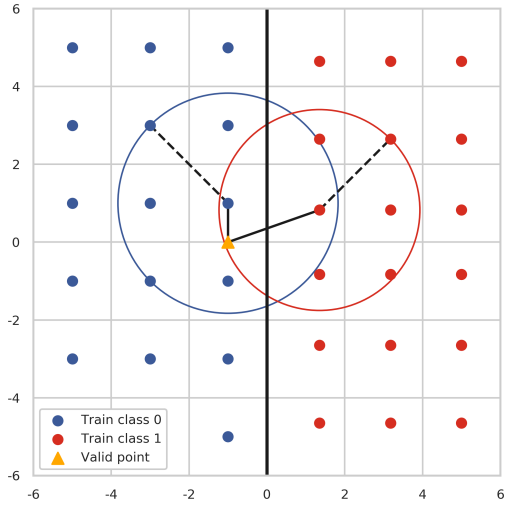

Figure 6.1: Example of a wrongly classified point query point.

$$\rho(g) \approx \sum_{i=0}^{k_{nn}} \omega_i e^{-\frac{||g-g^{(i)}||_{\ell 2}^2}{2\varepsilon_i^2}}. \tag{6.6}$$

Let $\overline{\rho}_i$ denotes the density at the sample $g^{(i)}$ obtained by the classical KNN approximation. The weights ω_i are computed as the result of the following optimisation problem:

$$\rho_{app}(g) = \sum_{i=0}^{k_{nn}} \omega_i e^{-\frac{||g-g^{(i)}||_{\ell 2}^2}{2\varepsilon_i^2}},$$

$$\mathcal{L}(\omega,\lambda) = \frac{1}{2}\sum_{i=1}^{k_{nn}} |\omega_i - \overline{\rho}_i|^2 + \lambda\left(\frac{k_{nn}}{N} - \int_{\mathcal{B}} \rho_{app}\, dg\right),$$

$$(\omega_*,\lambda_*) = \arg\inf_{\omega}\sup_{\lambda} \mathcal{L}(\omega,\lambda).$$

The interpretation is simple: the weights are close to the classical KNN estimated density (the Gaussian kernel being equal to one when evaluated at the sample), and when integrated on the ball, the approximation of the density retrieves the expected value of the mass in the ball. Let:

$$I_i = \int_{\mathcal{B}} e^{-\frac{||g-g^{(i)}||_{\ell 2}^2}{2\varepsilon_i^2}} dg.$$

The solution reads:

$$\omega_i^* = \overline{\rho}(g^{(i)}) + I_i \frac{k_{nn}/N - \sum_{j=0}^{k_{nn}} \overline{\rho}(g^{(j)}) I_j}{\sum_{j=0}^{k_{nn}} I_j^2}. \tag{6.7}$$

The following Example aims at illustrating the effect of the above introduced approximation on a classification task. Let $\Omega = [-5,5]^2$ be the domain, and $g = (g_0,g_1) \in \Omega$. We define the two classes as follows:

$$y = \begin{cases} 0, & g_0 > 0 \\ 1, & g_0 \leq 0 \end{cases}.$$

The sample size for the training set is $N_{0,1} = 18$. For each class the training set is uniformly distributed but with a different density (the density is higher for the class 1 as shown in Figure 6.1). The validation set is generated using a regular square mesh of Ω (with steps $\Delta g_0 = \Delta g_1 = 0.1$) where each node is a sample (it results in a validation sample size of $N_{0,1}^{te} = 5000$ for each class).

Figure 6.2 shows the result when the density is estimated via the classical KNN method and with the proposed Gaussian kernel correction. In this test, the accuracy is

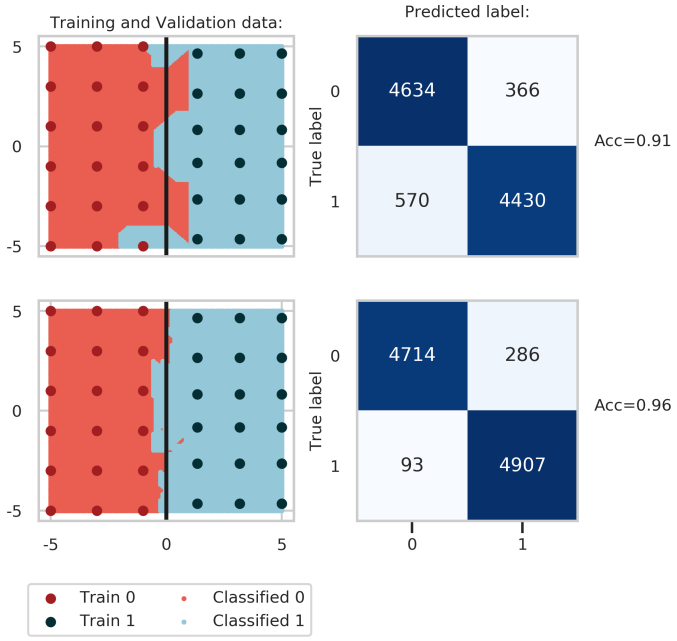

Figure 6.2: Comparison of the two methods in a binary classification example. Number of neighbours: 5. Upper: usual KNN method. Lower: RBF based approximation. Left: training and validation sets. Right: corresponding confusion matrices.

significantly increased using the proposed technique (we pass from 0.86 to 0.96).

6.3.2 Computing the Hausdorff distance of sets

One of the key steps of the proposed method is the approximation of the Hausdorff distance and the largest ball contained in the set $M^{(n)}$. Given the sets $S^n_{0,1}$, we can identify the $N_M \in \mathbb{N}^*$ samples, belonging to the validation set, which are in $M^{(n)} = \left(S_0^{(n)} \cap S_1^*\right) \cup \left(S_1^{(n)} \cap S_0^*\right)$. We denote $I_M^{(n)} \in \mathbb{N}$ the indices of these samples: $I_M^{(n)} = \left\{i \in 1, \dots, N_v \text{ such that } g^{(i)} \in M^{(n)}\right\}$. The pairwise distance between every element of $M^{(n)}$ is computed, and the pair of elements maximising the distance is chosen:

$$i_*, j_* = \arg \max_{i,j \in I_M^{(n)}} ||g^{(i)} - g^{(j)}||_{\ell^{p,n_g}}.$$

We then consider the segment relying the samples $g^{(i_*)}$ and $g^{(j_*)}$. The elements of this are characterised by the following expression. Let $\alpha \in$ [0,1] and the points: $g(\alpha) = (1-\alpha)g^{(i_*)} + \alpha g^{(j_*)}$. If the centre of the balls is chosen among the points of the segment, the problem reduces to finding α such that the radius of the ball inscribed in $M^{(n)}$ is the largest:

$$\alpha_* = \arg \sup_{\alpha \in [0,1]} \varepsilon,$$
$$\mathcal{B}(g(\alpha), \varepsilon) \subseteq M^{(n)}.$$

This problem is solved numerically by extensive search: the segment is discretised by considering a number of points on it, where the evaluation of the ball radius is performed.

Remark 11
During the enrichment process, it might happen that there are no elements in the reservoir belonging to the ball chosen to reduce the Hausdorff distance between the sets. We propose to add to the augmented set the centre of the ball, labelled as the closest sample belonging to the validation set.

6.3.3 Summary of the method

The overall method is summarised hereafter. Two validation sets are given, namely $S^*_{0,1} \subset \Omega$, in the form of sets of validation samples $g^{(v)}$. At the beginning of the procedure, we have two augmented sets $S^{(0)}_{0,1} \subset \Omega$, given in form of sets of samples $g^{(0)}$. At the beginning of a generic iteration of the method, say n, we have two augmented sets $S^{(n)}_{0,1}$.

1. *Evaluate the intersections between the validation sets and the current augmented sets:*$M^{(n)} = (S^*_0 \cap S^{(n)}_1) \cup (S^*_1 \cap S^{(n)}_0)$. To do so:
   (a) Evaluate the densities $\rho^{(n)}_{0,1}$ in the validation sample points $g^{(v)}$ by using the method described in Section 6.3.1.

(b) Perform a Bayesian classification providing the labels y.

(c) Compare the labels with the true validation labels y^*.

(d) If $y \neq y^*$ then $g^{(v)} \in M^{(n)}$.

2. We compute an approximation of the Hausdorff distance, by evaluating the maximum of the distance between the well-classified validation samples and the wrongly classified ones, that belong to $M^{(n)}$.

3. We compute the largest ball that is contained in $M^{(n)}$, by following the steps presented in Section 6.3.2.

4. We compute $S_{0,1}^{(n+1)}$ by adding to them the elements of the reservoir which are contained in the largest ball computed at the previous step, following Equation (6.3).

Remark 12
The choice of a Bayesian classification derives naturally from the distribution mixture hypothesis given in Equation (6.1). In particular, we have: $\mathbb{P}(Y = i|G = g) > \frac{1}{2} \Longleftrightarrow \pi_i \rho_i(g) > \pi_j \rho_j(g)$, $i,j \in \{0,1\}, i \neq j$. *Where i is a realisation of the random variable Y defined on $(\Omega_{cl} = \{0,1\}, \mathcal{A}_{cl}, \mathbb{P})$ (corresponding to the class) and g a realisation of G defined on $(\Omega \subseteq \mathbb{R}^{n_g}, \mathcal{A}, \mathbb{P})$ (corresponding to the observation).*

The pseudo-code of the method is given in Algorithm 2.

6.4 Numerical experiments

In this section, several numerical experiments are proposed to illustrate the enrichment method.

6.4.1 Two-dimensional cases

A two-dimensional application is performed on three study cases for which we consider $\Omega = [0,1]^2$. For each study case, we randomly generated 2000 samples following a uniform law over Ω. The first half is gathered into the validation set, whereas the second half is gathered into the test set. Figure 6.4 shows the validation set for each study case. The colour corresponds to the label and the black line corresponds to the true delimitation of the two classes.

The same uniform random process was performed to construct the initial augmented set (of size 20) and the reservoir of simulation (of size 1000). A summary of the sets is given below:

- Input of the algorithm: validation set (of size 1000), test set (of size 1000), initial augmented set (of size 20) and reservoir of simulations (of size 1000). Each sample (in $\Omega = [0,1]^2$) is an observation (input of the classifier) with its corresponding label (output of the classifier).

Algorithm 2 Augmented set Construction (ASE-HD): Overall algorithm.

Require: Tr;V;R;k_{nn}
{Input: Initial Augmented set; Validation set; Reservoir; Number of neighbours.}
Require: $\pi \leftarrow (\pi_0,\pi_1)$ {*A priori* in the binary classification case.}
$\mu_v \leftarrow 0$ {Initialise the score on the validation set.}
while $\mu_v < 1$ **do**
$\mu_v, V^{cl} \leftarrow classify(Tr, V, k_{nn}, \pi)$
{Classify the Validation set from the Augmented set. See Algorithm 3.}
$pnt_0 \leftarrow Hausd(V_0, V_1^{cl_0})$
{Get points maximising the Hausdorff distance (for class 0)†. See Algorithm 4.}
$pnt_1 \leftarrow Hausd(V_1, V_0^{cl_1})$ {Same for class 1†. See Algorithm 4.}
$r_0, c_0 \leftarrow computeBall(V_0, V_1^{cl_1}, pnt_0)$
{Get the centre and radius of the biggest ball in the intersection on the segment delimited by pnt_0. See Algorithm 5.}
$r_1, c_1 \leftarrow computeBall(V_1, V_0^{cl_0}, pnt_1)$ {Same for class 1. See Algorithm 5.}
for each class i **do**
$nbElem \leftarrow 0$ {Initialise the number of added elements.}
for $g_R \in R_i$ **do**
if $g_R \in \mathcal{B}(c_i, r_i)$ **then**
$Tr \leftarrow Tr \cup g_R$ {We add the element to the current Augmented set.}
$nbElem \leftarrow nbElem + 1$ {Increment the number of added elements.}
end if
end for
if $nbElem == 0$ **then**
$Tr = Tr \cup c_i$ {We add the centre of the ball to the current Augmented set.}
end if
end for
end while
return Tr {Output of the algorithm.}

†: $V_i^{cl_j}$ are elements of the validation set belonging to class i and labelled j. These steps allow the computation of $M^{(n)}$ described above in the paper.

- Output of the algorithm: augmented set and classification scores on the validation and test set.

In this study we assume that the reservoir is unbiased. The number of nearest neighbours is set to $k_{nn} = 5$.

Figure 6.5 shows the constructed training set (augmented set once the algorithm has stopped) samples for each study case.

Two main points are highlighted by this figure:

- The whole initial database is not a must-have, only a small fraction of it is actually useful in view of improving the classification score.

Algorithm 3 Classify method: Classify samples of a Validation set from a Training set.

Require: Tr;V;k_{nn};π {Input: Training set; Validation set; Number of neighbours; *A priori.*}
  $\mu_v \leftarrow 0$ {Initialise the score on the validation set.}
  $V^{cl} \leftarrow \emptyset$ {Initialise the output (classified samples).}
  for each class i **do**
    $j \leftarrow \{0,1\} \setminus i$ {In the binary classification case.}
    for $g_v \in V_i$ **do**
      $\rho_i(g_v) \leftarrow estimateDensity(Tr_i, k_{nn}, g_v)$ {Estimate density of g_v from $Tr_i^{\dagger}$.}
      $\rho_j(g_v) \leftarrow estimateDensity(Tr_j, k_{nn}, g_v)$ {Same for the other class†.}
      if $\pi_i\rho_i > \pi_j\rho_j$ **then**
        $\mu_v \leftarrow \mu_v + 1$ {Increase the score if well classified.}
        $V_i^{cl_i} \leftarrow V_i^{cl_i} \cup g_v$ {Add the element belonging to class i and labelled i.}
      else
        $V_i^{cl_j} \leftarrow V_i^{cl_j} \cup g_v$ {Add the element belonging to class i and labelled j.}
      end if
    end for
  end for
  $\mu_v \leftarrow \mu_v / \#V$ {Renormalise the score.}
  return μ_v;V^{cl} {Output of the algorithm.}

†: Using the method described in Section 6.3.1. See Equation (6.6)-(6.7).

- The selected samples to construct the augmented set are mainly closed to the class delimitation.

Figure 6.6 shows the scores on the test sets for each study case. As the algorithm is performed on the validation set, the score on the validation set is 1 by construction (which then induces an overfitting). Despite this overfitting, the constructed augmented set ensures a high score on the test set for the three study cases. This is particularly true for the first and second study cases with an average success rate higher than 0.97.

6.4.1.1 Influence of the KNN parameter

In this section, we empirically discuss about the KNN free parameter. Study case 0 was run with 2, 3, 5 and 7 neighbours in the algorithm. The three quantities considered for the comparison are:

- The score.
- The computation time (normalised with respect to the longest one).
- The compression (augmented set size normalised with respect to the reservoir size.)

Algorithm 4 Hausd method: Return the two samples maximising the Hausdorff distance.

Require: V_i;$V_j^{cl_i}$ {Input: Validation set restricted to class i; Validation set restricted to class j and labelled i.}
  $pnt \leftarrow \emptyset$ {Points maximising the Hausdorff distance.}
  if $\#V_j^{cl_i} > 0$ **then**
    $d_A \leftarrow closest(V_j^{cl_i}, V_i)$ {For each element of $V_j^{cl_i}$ return the distance of its closest neighbour in $V_i^{\dagger}$.}
    $p_A \leftarrow argmax(d_A)$ {Maximum distance position.}
    $pnt_A = (V_i(p_A), V_j^{cl_i}(p_A))$ {Samples maximising the distance.}
    $d_B \leftarrow closest(V_i, V_j^{cl_i})$ {For each element of V_i return the distance of its closest neighbour in $V_j^{cl_i}$ †.}
    $p_B \leftarrow argmax(d_B)$ {Maximum distance position.}
    $pnt_B = (V_i(p_B), V_j^{cl_i}(p_B))$ {Samples maximising the distance.}
    if $d_A > d_B$ **then**
      $pnt \leftarrow pnt_A$ {Then, d_A is the Hausdorff distance.}
    else
      $pnt \leftarrow pnt_B$ {Then, d_B is the Hausdorff distance.}
    end if
  end if
  return pnt {Output of the algorithm.}

†: In this paper, we consider the ℓ^{∞} distance.

Results are given in Figure 6.7. For each number of neighbours, the random part (pruning) was repeated 10 times.

Similar results for study cases 1 and 2 are shown in Figures 6.13 and 6.14 respectively in the Appendix.

Despite the score is significantly high for the three study cases, it globally increases as the number of neighbours decreases. Moreover, a lower number of neighbours induces a lower computation time and higher compression. Computation time and compression are highly correlated with a Pearson correlation close to 0.97 (0.98 for study case 1 and 0.97 for study case 2).

6.4.2 A model in electrophysiology of cells

This part is devoted to an example in electrophysiology. The observed model output, called action potential (AP) is the potential difference across the cell membrane. This is influenced by the value of several parameters which represent the conductances of some of the ion channels of the cell. The model we consider is called Minimal Ventricular (MV), presented in [BOCF08]; it is a system of parametric ordinary differential equations. We focus on three classification problems: given the model output determine if the conductances of sodium, calcium and potassium are above or below a certain threshold.

Algorithm 5 computeBall method: Return the ball to consider for enrichment.

Require: V_i;$V_j^{cl_j}$; pnt_i {Input: Validation set of class i; Validation set of class j classified j; Hausdorff points.}
$r \leftarrow \emptyset$; $c \leftarrow \emptyset$ {Initialise the radius and centre of the ball.}
$line \leftarrow lineFrom(pnt_i)$ {Extract discretised line between the two Hausdorff points.}
$d_i \leftarrow dist(line, V_i)$ {For each element of $line$, compute closest distance to V_0.}
$d_j \leftarrow dist(line, V_j^{cl_j})$ {For each element of $line$, compute closest distance to the well-classified samples in the other class.}
$\widehat{d_i}$;$\widehat{d_j} \leftarrow reorder(d_i,d_j)$ {Reorder distances with respect to the descending order of d_i.}
$cpt \leftarrow 0$
while $d_i(cpt) > d_j(cpt)$ **do**
  $cpt \leftarrow cpt + 1$ {Go closer to the point belonging to V_i.}
end while
$r \leftarrow d_i(cpt)$ {Actualise the radius†.}
$c \leftarrow line(cpt)$
return r;c {Output of the algorithm.}

†: For sake of clarity a scheme is given in Figure 6.3.

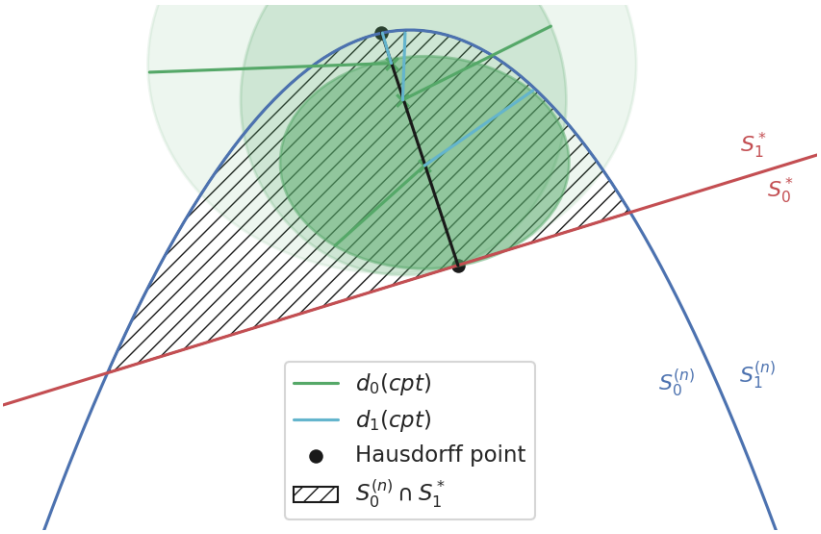

Figure 6.3: Scheme for Algorithm 5. We call Hausdorff points the two points for which the Hausdorff distance is computed. $S_0^{(n)} \cap S_1^*$ corresponds to the area where the samples of the validation set belonging to class 1 are labelled 0 at step n (i.e. belonging to $S_0^{(n)}$). We move on the segment delimited by the Hausdorff points, starting from the farthest one from S_0^*. At each step, we compute the distances to S_0^* ($d_0(cpt)$) and $S_1^{(n)}$ ($d_1(cpt)$).

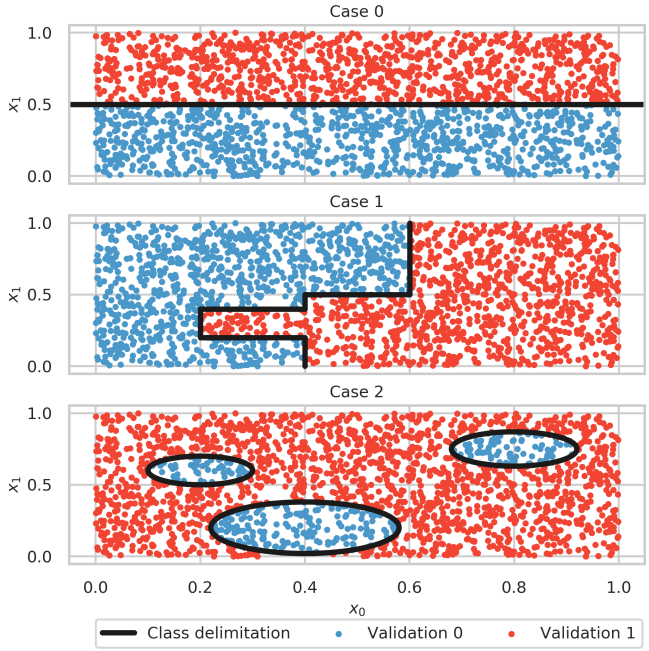

Figure 6.4: Study cases.

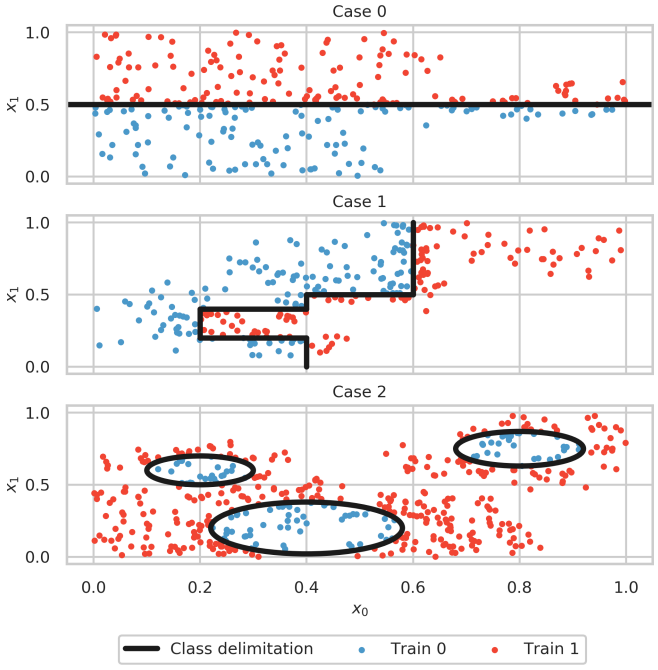

Figure 6.5: Constructed training sets (augmented sets once the algorithm stops).

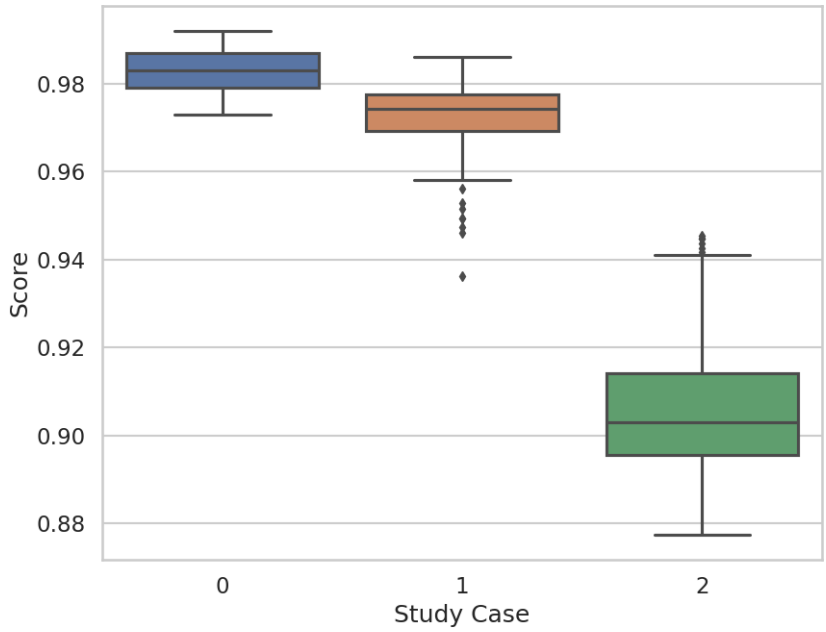

Figure 6.6: Boxplots of the scores obtained on the test set for each study case.

The dataset is synthetic and the numerical method used to approximate the model solution is a third order Backward Differentiation Formula (BDF3) with a time step $\Delta t = 0.1$ms. A periodic source term in the equation is repeated every 1200ms and its parametrisation is given in Table 6.1.

Duration (ms)	Amplitude (pA/pF)
4.0	0.1

Table 6.1: Stimuli parameters.

By starting from the third stimulation the system reaches periodicity (the ℓ^2 norm of the difference between two consecutive periods varies by less than 10^{-3}) we decided to only store the third period for this study.

A total of $n_s = 2420$ signals were generated with random triplets conductances (for sodium, calcium and potassium) following a uniform law over $[0.6,1]^3$. It follows that for a realisation $x = [x_{sodium}, x_{calcium}, x_{potassium}]$, the component x_i means that channel i is blocked at $100 * (1 - x_c)\%$. We consider the control case (as a reference) for the realisation $x = [1,1,1]$ which leads to 100% of activity for each channel.

For each component c of a realisation x, the labels y_c are given by:

$$y_c = \begin{cases} 0 \text{ if } x_c < 0.8 \text{ ("blocked")} \\ 1 \text{ otherwise ("not blocked")} \end{cases} . \tag{6.8}$$

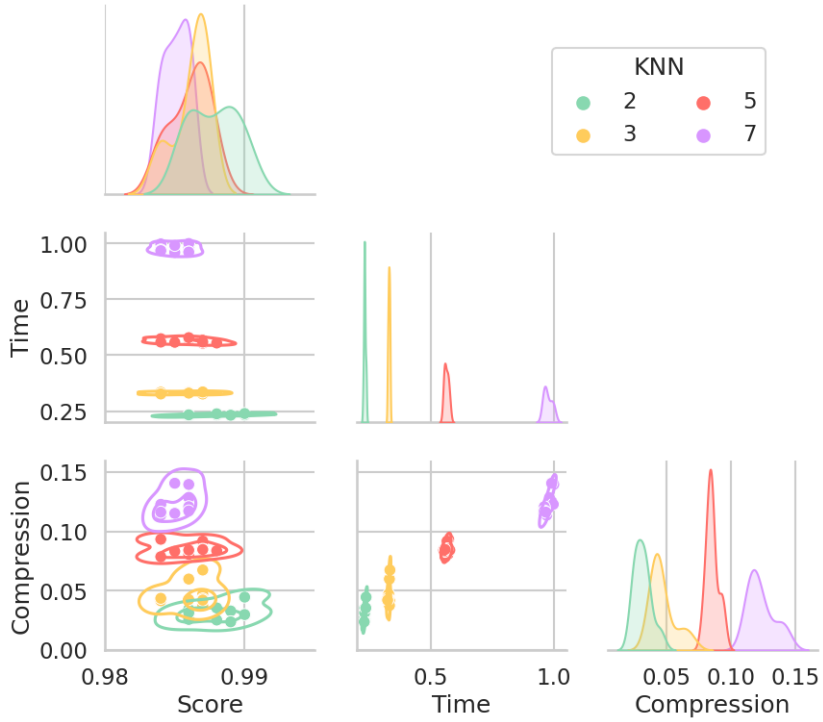

Figure 6.7: Influence of the number of neighbours on the score, the computation time and the compression for the study case 0.

The value 0.8 corresponds to the conductance threshold for the classification task described at the beginning of this section. As we have three parameters, we divided the problem into three classification tasks: sodium, calcium and potassium conductances classification. An example of AP signals at control case ($x = [1, 1, 1]$) and in random case is shown in Figure 6.15 of the Appendix.

6.4.2.1 Biased data

Different biased datasets were generated from these $n_s = 2420$ simulated APs. These biased signals were obtained by computing the Fourier transform and putting to zero the entries corresponding to the higher frequencies. We considered three different levels of bias (expressed in terms of energy) as presented in Table 6.2.

Bias level	Relative ℓ^2 error norm
Low	0.020
Medium	0.035
High	0.065

Table 6.2: Biased datasets.

An example of an AP signal with its different levels of bias is shown in Figure 6.8.

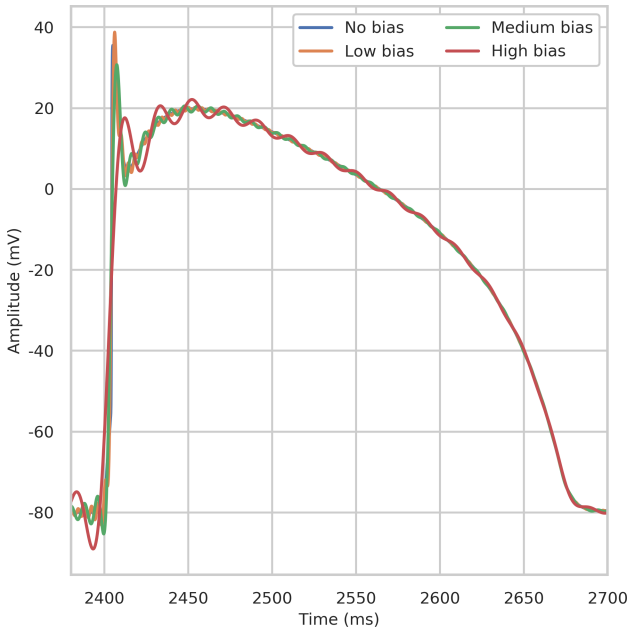

Figure 6.8: Sample of an action potential signal generated by the MV model with its different levels of bias.

6.4.2.2 Dictionary entry computation

For each sample (AP signal), we consider $n_g = 24$ observable quantities. These correspond to pair times and amplitudes in different phases of the AP signal. Its computations are the following:

- The first two dictionary entries are the amplitude and time of the depolarization peak.
- The next two dictionary entries are the amplitude and time of the notch.
- Then the amplitude and time of the plateau are considered.
- The others quantities are the amplitudes and times of the different percentage X of repolarization with respect to the plateau amplitude: where X is equal to 10, 20, 30, 40, 50, 60, 70, 80 and 90. It corresponds to a variant of the action potential duration at $X\%$ of repolarization (APDX), where the reference is the plateau amplitude instead of the depolarization amplitude.

They are computed in the same way for each sample. An example of extracted quantities is shown in Figure 6.9.

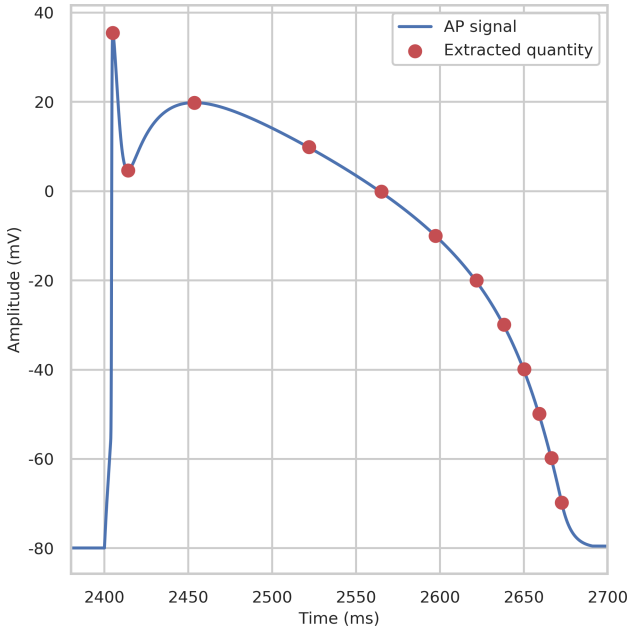

Figure 6.9: Sample of an action potential signal generated by the MV model (control case: $x = [1, 1, 1]$) with the extracted quantities to generate the dictionary entries.

We denote $g_i^{(j)}$ the i^{th} dictionary entry of the j^{th} AP signal. Considering the control case as a reference, we propose to consider the following translated dictionary entries:

$$g_i^{(j)} = g_i^{(j)} - g_i^{(ctrl)}, \forall i,j.$$

It follows that, in the control case, we have $g_i^{(ctrl)} = 0, \forall i = 1, \ldots, n_g$. All the samples were then transformed in such a way that the compact domain Ω is the hypercube of dimension $n_g = 24$, side 1 and centred at $c = (\frac{1}{2}, \ldots, \frac{1}{2}) \in \mathbb{R}^{n_g}$. Inputs and outputs of the model are summarised below:

- Input of the model to generate one sample: $x = [x_{sodium}, x_{calcium}, x_{potassium}] \in [0.6, 1]^3$.
- Output of the model: computed entries rescaled with respect to the control case (computed entries for x_c = [1,1,1]) and its corresponding label from x given by Equation (6.8).

6.4.2.3 Datasets preprocessing

Two study cases are performed: in the first one, we assume that the validation set covers Ω whereas in the second one we consider an incomplete validation (the validation set covers only a subset of Ω). To do so, from the unbiased dataset, we randomly extract $n_v = 89$ from the $n_s = 2420$ signals in such a way that 84 of them have a sodium and

calcium activity higher than 0.85. The 5 others are randomly chosen in such a way that at least one sample belongs to the other class (sodium and/or calcium conductance is lower than the threshold). Dataset's sizes are summarised in Table 6.3.

Validation set	n_t (test)	n_v (validation)	n_{tr} (initial augmented)	n_d (database)
Complete: Covers Ω	1000	400	20	1000
Incomplete: Partially covers Ω	1000	89	20	1000

Table 6.3: Datasets sizes.

Test, validation and initial augmented sets are randomly extracted from the whole unbiased dataset ($n_s = 2420$). The database can be biased or unbiased depending on the study (chosen samples are the same, but with different biases). The random process is performed in such a way that a selected sample belongs to only one set and cannot be selected more than once. Figure 6.10 shows the densities of the variable x for the validation and test sets (for each class), in the sodium classification task.

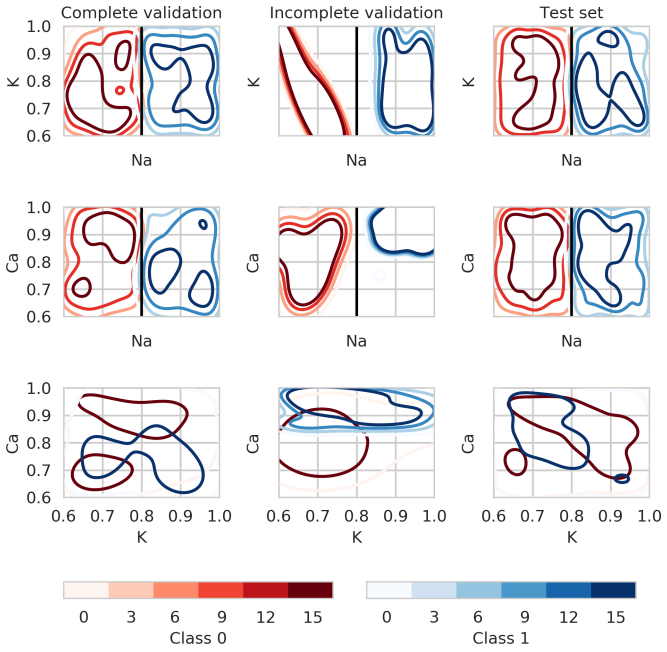

Figure 6.10: Densities of validation and test sets for sodium classification. Black lines correspond to the class delimitation.

As we can see, when the complete validation case is considered, the density of x is almost uniform over the whole domain of x (meaning that we have samples for almost all possible values of x). On the contrary, for the incomplete validation case (in the centre) we clearly see that there are regions of the domain of x in which we do not have samples.

6.4.2.4 Computational results

All the following results were obtained using $k_{nn} = 5$ nearest neighbours.

Comparison between complete and incomplete validation set

Figure 6.11 shows the scores obtained with a complete and incomplete validation set.

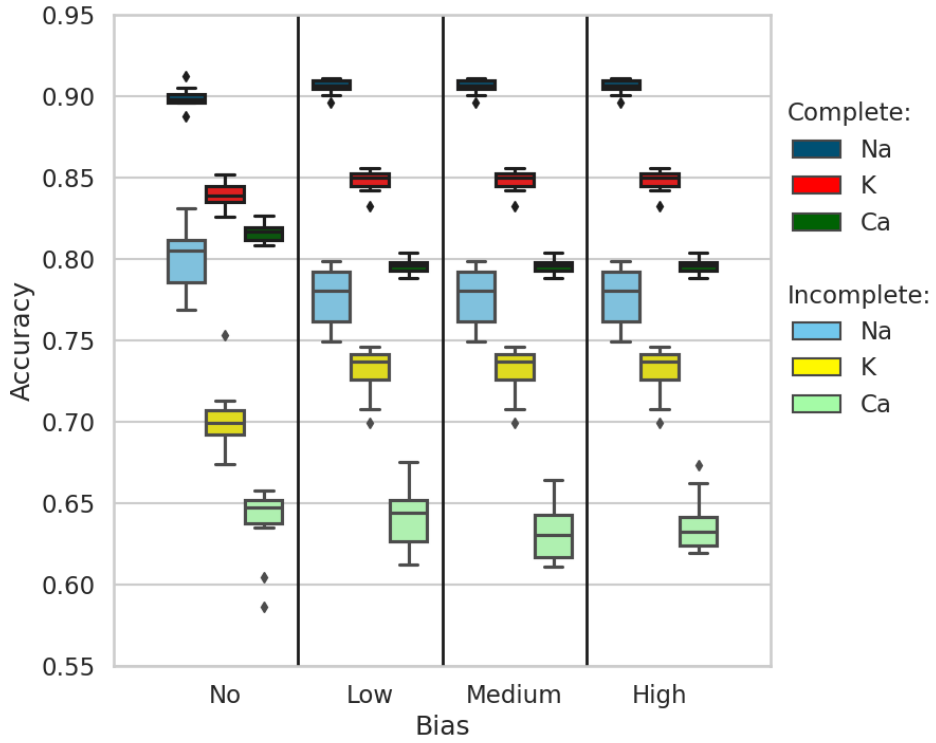

Figure 6.11: Scores obtained with a complete and incomplete validation set.

1. Complete validation set:
   (a) The sodium conductance is easy to classify, whereas calcium conductance is the most difficult to infer. The fact that potassium and calcium conductances are more difficult to classify is due to the compensation effect between these two channels (see Figure 6.15), which is a known phenomenon in electrophysiology.
   (b) The scores are not significantly impacted by the bias as the proposed method naturally rejects it.
2. Incomplete validation set:
   (a) The calcium conductance classification shows the lowest success rate whereas the potassium conductance classification shows the highest score. The fact

that the potassium has the highest score is expected as no data were removed for this case.

(b) The score obtained in the unbiased case for the sodium is close to its expected approximation: around 81% (see Section 6.6.1 of the Appendix for more details).

3. Complete vs Incomplete validation set:

(a) The test score is lower in the incomplete validation set case. This is because there are regions of Ω in which we do not have samples of the dataset. As we do not have information in these empty regions, the score is lower.

(b) For the same reasons as above, the variability on the test score is higher when the validation set is incomplete.

A comparison with the construction of a classifier considering the full reservoir of data as the training set is given in Table 6.4. The same conditions were considered for the three methods (ASE-HD, SVM and KNN). Indeed, for the 'No Bias' scenario we put in the reservoir unbiased samples, for the 'Low' scenario we considered only samples with a low level bias in the reservoir, and we proceed analogously for the other scenarios. For all the cases, the samples of the test set are unbiased, meaning, they are drawn from the 'true' system. In particular, in absence of bias, considering the whole reservoir as the training set is globally better. However, in the presence of bias, the augmented set construction method proposed is better. Moreover, the construction method allows to get a similar classification success rate irrespective of bias. This is due to the method itself which reject biased data in an automated way.

Remark 13
In the KNN algorithm implemented in Scikit-Learn [PVG+11], we consider the k_{nn}^{th} closest samples (from the training set) of a query point irrespective of the class they belong to. We then classify the query point using the majority vote strategy. This method is quite different of the proposed strategy proposed in this paper. In particular, we consider the k_{nn}^{th} closest samples from the training set of a query point for each class to estimate the density over the two classes (for a binary classification). We then consider a Bayesian approach to classify the query point. This could justify the success rate difference between the No Bias case in the ASE-HD method and KNN method.

Remark 14
In this paper we considered a Bayesian approach to classify a query point. However, the augmented set construction method is not restricted to a particular classification method (nor density approximation).

Database and validation set enrichment

Study/Biased Reservoir	ASE-HD*	SVM	KNN
Sodium			
No Bias	0.91	0.94	0.94
Low	0.91	0.81	0.82
Medium	0.91	0.81	0.80
High	0.91	0.54	0.57
Average (std)	0.91 (0)	0.78 (0.17)	0.78 (0.15)
Potassium			
No Bias	0.83	0.91	0.89
Low	0.84	0.86	0.76
Medium	0.84	0.82	0.83
High	0.84	0.83	0.80
Average (std)	0.84 (0.01)	0.86 (0.04)	0.82 (0.05)
Calcium			
No Bias	0.81	0.86	0.87
Low	0.81	0.60	0.67
Medium	0.79	0.78	0.79
High	0.80	0.68	0.64
Average (std)	0.80 (0.01)	0.73 (0.11)	0.74 (0.11)
*See Figure 6.11, left panel and blue legend.			

Table 6.4: Comparison between the augmented set construction method and common classification techniques (using Scikit-Learn library [PVG+11] with default parameters) considering the whole reservoir (biased or unbiased samples depending on the scenario) as the training set. Values correspond to the classification success rate on the test set which is unbiased.

As described in Section 6.2.4.2, once the augmented set enrichment process is performed on the incomplete validation set, we enrich the validation set with data from the database. In the case where we have a bias, we may exploit some statistical information on the bias to generate more pertinent labelled samples. We recall that we have 4 different study cases based on the database (see Section 6.4.2.1): without bias and with a low, medium and high level of bias. We assume that we know the *a priori* for the two classes: $\pi_0 = \pi_1 = \frac{1}{2}$. Then, we enriched the validation set in such a way the number of samples is each class is the same, with $n_v = 400$ (we added 311 samples). See Table 6.3.

Unbiased case In the unbiased case, we compute the dictionary entry mean and standard deviation for each class of the incomplete validation set. We denote $\widehat{\pi}_i$ the estimated *a priori*. Then, we randomly brows each sample of the database (for each class). While $n_v < 400$, if one of the entries is outside the corresponding (i.e same class) mean plus/minus the standard deviation, we add it to the validation set (and remove it from the database) if the following equation holds:

$$\min_i \widehat{\pi}_i^{(n+1)} > \min_i \widehat{\pi}_i^{(n)},$$

with $\widehat{\pi}_i^{(n+1)}$ the *a priori* computed considering the sample into the validation set and $\widehat{\pi}_i^{(n)}$ the *a priori* computed before considering the current sample into the validation set. In other words, it aims to consider the assumptions on the true *a priori* π_i described above.

Biased case For the biased case, we compute the average and standard deviation difference (in the dictionary entry space) between the incomplete validation set and the simulated data with the same parameters:

$$\begin{cases} b_m = \mathbb{E}\Big(D_{\theta_v} - V_{\theta_v}\Big) \\ b_s = \sqrt{\mathbb{E}\Big((D_{\theta_v} - V_{\theta_v})^2\Big)} \end{cases},$$

with $b_j \in \mathbb{R}^{n_g}$ the mean ($j = m$) or the standard deviation ($j = s$) and where V_{θ_v} is the incomplete validation set and D_{θ_v} is the simulated dataset obtained with θ_v as parameter entries of the simulated model. Then, from these statistics, for each sample of the database, we generate 4 ghosts samples following the approach described in Section 6.2.4.2. Here, we assume that the bias computed on the validation set is preserved on the empty region.

Results The results are shown in Figure 6.12.

1. The enrichment scenario always induces a higher classification success rate on the test set. It also reduces the variability.
2. Whithout bias, the classification success rate is close to the complete validation set scenario.
3. The sodium channel is the easiest to classify if compared with the potassium and calcium channels.
4. The highest gain with the enrichment is for the calcium channel.
5. The gain with the enrichment for the sodium channel is the lowest. This is because the part of signal induced by the sodium channel (depolarisation) is particularly sensitive to the bias (see Figure 6.15).

6.5 Conclusion

In the present work a method is proposed to enrich available experimental datasets by using numerical simulations in view of improving classification tasks performances. This is an example of potential interaction between statistical learning and mathematical

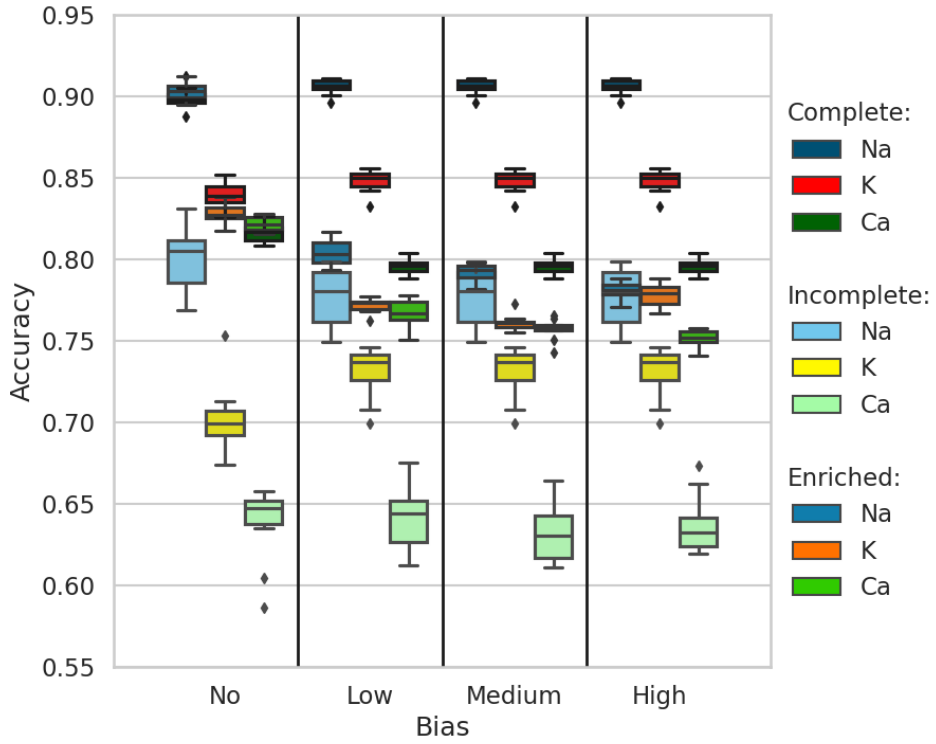

Figure 6.12: Scores on the test set considering incomplete/complete validation set and enriched validation set.

modelling. The method is based on the probabilistic description of the observations of a phenomenon and a characterisation of the classification performances based on set distances. The main properties of the method have been investigated from a theoretical point of view and illustrated through some numerical experiments. The systematic construction and enrichment of the augmented set can have a significant impact on the classification score. The proposed method performs a bias rejection to some extent, and, if statistical information on a model bias is available, these can be naturally integrated in the algorithm.

6.6 Appendix

Proof of Lemma 1.

By definition of the symmetric difference, we have:

$$S_0^* \Delta S_0^{(n)} = (S_0^* \setminus S_0^{(n)}) \cup (S_0^{(n)} \setminus S_0^*).$$

$$\Longleftrightarrow$$

$$S_0^* \Delta S_0^{(n)} = (S_0^* \cap S_0^{(n)^C}) \cup (S_0^{(n)} \cap S_0^{*^C}),$$

where $S_0^{(n)^C} = \Omega \setminus S_0^{(n)}$ and $S_0^{*^C} = \Omega \setminus S_0^*$ are the complementary sets of $S_0^{(n)}$ and S_0^* respectively. It follows that:

$$S_0^* \Delta S_0^{(n)} = (S_0^* \cap S_1^{(n)}) \cup (S_0^{(n)} \cap S_1^*) = M^{(n)}.$$

The proof for $S_1^* \Delta S_1^{(n)}$ is similar. ∎

Proof of Proposition 7.

By definition of S_j^* and $S_j^{(n)}$ (see Equation (6.2)), we have:

$$(S_0^* \cap S_1^{(n)}) \cap (S_1^* \cap S_0^{(n)}) = \emptyset.$$

Then, $M^{(n)}$ is a disjoint union of two sets. This implies that:

$$\mu_L(M^{(n)}) = \mu_L(S_0^* \cap S_1^{(n)}) + \mu_L(S_1^* \cap S_0^{(n)}).$$

Remark that, by definition of the Lebesgue measure on a set and due to the compactness of the sets, we have the following inequalities:

$$0 \leq \mu_L(M^{(n)}) < +\infty.$$

It is straightforward to show that:

$$\mu_L(M^{(n)}) = 0 \Longleftrightarrow \mu_v^{(n)} = \mu_* \text{ almost surely,}$$

Let assume that $\mu_L(M^{(n)}) > 0$. It follows that at least one of the following inequalities is satisfied:

$$\begin{cases} \mu_L(S_0^* \cap S_1^{(n)}) > 0 \\ \mu_L(S_1^* \cap S_0^{(n)}) > 0 \end{cases}.$$

Let S' be the set such that:

$$S' = \arg\max \Big(\mu_L(S_0^* \cap S_1^{(n)}), \mu_L(S_1^* \cap S_0^{(n)}) \Big).$$

We then have $\mu_L(S') > 0$. Therefore, $\exists g_{n+1} \in S'$ and $\varepsilon > 0$such that the ball $\mathcal{B}(g_{n+1}, \varepsilon) \subseteq S'$. By definition of $M^{(n)}$ (see Section 6.2.2), we have:

$$M^{(n+1)} = M^{(n)} \setminus \mathcal{B}.$$

As $\mathcal{B} \in S' \subseteq M^{(n)}$ and $\mu_L(\mathcal{B}) > 0$, we have:

$$0 \leq \mu_L(M^{(n+1)}) < \mu_L(M^{(n)}).$$

We have a sequence of measures which is strictly decreasing and bounded. Thus, this sequence converges to its minimum. Let assume that this minimum is $\delta > 0$. Then, it exists a non-empty ball such that the measure will decrease, which is impossible. It follows that:

$$\lim_{n \to +\infty} \mu_L(M^{(n)}) = 0.$$

Therefore,

$$\lim_{n \to +\infty} S_i^{(n)} = S_i^*,$$

almost everywhere for $i = 0$ or 1. Hence, almost surely, we have:

$$\lim_{n \to +\infty} \mu_v^{(n)} = \mu_*.$$

■

Proof of Corollary 2.

By definition, $\forall n \in \mathbb{N}$, we have:

$$\mu_v^{(n)} = \int_{S_0^{(n)}} \pi_0 \rho_0^v dg + \int_{S_1^{(n)}} \pi_1 \rho_1^v dg.$$

Then, at iteration $n+1$, we have:

$$\mu_v^{(n+1)} = \int_{S_0^{(n+1)}} \pi_0 \rho_0^v dg + \int_{S_1^{(n+1)}} \pi_1 \rho_1^v dg,$$

with:

$$S_0^{(n+1)} = \begin{cases} S_0^{(n)} \cup \mathcal{B}_* \text{ if } \mathcal{B}_* \subseteq S_0^* \cap S_1^{(n)} \\ S_0^{(n)} \setminus \mathcal{B}_* \text{ if } \mathcal{B}_* \subseteq S_1^* \cap S_0^{(n)} \end{cases}.$$

Let us consider the first scenario: $S_0^{(n+1)} = S_0^{(n)} \cup \mathcal{B}_*$. Then using the fact that the sets are disjoint, we have:

$$\mu_v^{(n+1)} = \int_{S_0^{(n)}} \pi_0 \rho_0^v dg + \int_{\mathcal{B}_*} \pi_0 \rho_0^v dg + \int_{S_1^{(n)}} \pi_1 \rho_1^v dg - \int_{\mathcal{B}_*} \pi_1 \rho_1^v dg,$$

which immediately yields to:

$$\mu_v^{(n+1)} - \mu_v^{(n)} = \int_{\mathcal{B}_*} (\pi_0 \rho_0^v - \pi_1 \rho_1^v) dg \geq 0.$$

Here, we assumed that $\mathcal{B}_* \subseteq S_0^* \cap S_1^{(n)}$. The inequality is given by the definition of S_0^*. On this set, we have: $\pi_0 \rho_0^v - \pi_1 \rho_1^v > 0$. The equality is then obtained if and only if $\mu_L(\mathcal{B}_*) = 0$. Considering the second scenario, we finally obtain:

$$\mu_v^{(n+1)} - \mu_v^{(n)} = \int_{\mathcal{B}_*} |\pi_0 \rho_0^v - \pi_1 \rho_1^v| dg \geq 0.$$

■

Proof of Lemma 2.

Let us focus on the first equality of the lemma (the proof for the second equality is similar). We have:

$$(S_0^{te} \cup b_0) \setminus b_1 = \Big(S_0^{te} \setminus b_1\Big) \cup \Big(b_0 \setminus b_1\Big)$$

As $b_1 \cap b_0 = \emptyset$ we have:

$$(S_0^{te} \cup b_0) \setminus b_1 = \Big(S_0^{te} \setminus b_1\Big) \cup b_0 = \Big(S_0^{te} \setminus (S_1^m \cap S_0^{te})\Big) \cup b_0.$$

$$\Longleftrightarrow$$

$$(S_0^{te} \cup b_0) \setminus b_1 = \Big(S_0^{te} \setminus S_1^m\Big) \cup b_0 = \Big(S_0^{te} \setminus S_1^m\Big) \cup \Big(S_0^m \cap S_1^{te}\Big)$$

$$\Longleftrightarrow$$

$$(S_0^{te} \cup b_0) \setminus b_1 = \Big(S_0^m \cap S_0^{te}\Big) \cup \Big(S_0^m \cap S_1^{te}\Big) = S_0^m \cap \Big(S_0^{te} \cup S_1^{te}\Big).$$

Since $S_0^{te} \cup S_1^{te} = \Omega$, we finally obtain:

$$(S_0^{te} \cup b_0) \setminus b_1 = S_0^m.$$

■

Proof of Proposition 8.

We have:

$$\mu_b = \int_{S_0^m} \pi_0 \rho_0^{te} dg + \int_{S_1^m} \pi_1 \rho_1^{te} dg.$$

Then from Lemma 2 and based on sets definition, we have:

$$\mu_b = \int_{S_0^{te}} \pi_0 \rho_0^{te} dg + \int_{b_0} \pi_0 \rho_0^{te} dg - \int_{b_1} \pi_0 \rho_0^{te} dg + \int_{S_1^{te}} \pi_1 \rho_1^{te} dg + \int_{b_1} \pi_1 \rho_1^{te} dg - \int_{b_0} \pi_1 \rho_1^{te} dg$$

$$\Longleftrightarrow$$

$$\mu_b = \mu_* + \int_{b_0} (\pi_0\rho_0^{te} - \pi_1\rho_1^{te})dg + \int_{b_1} (\pi_1\rho_1^{te} - \pi_0\rho_0^{te})dg.$$

By virtue of the definition of the sets b_0, b_1, it holds:

$$\begin{cases} g \in b_0 \Longrightarrow \pi_1\rho_1^{te} > \pi_0\rho_0^{te} \\ g \in b_1 \Longrightarrow \pi_0\rho_0^{te} > \pi_1\rho_1^{te} \end{cases}.$$

It immediately leads to $\mu_b \leq \mu_*$. Moreover,

$$\mu_b = \mu_* \Longrightarrow \mu_L(b_i) = 0, \ (i \in \{0,1\}),$$

and,

$$\mu_L(b_i) = 0, \ (i \in \{0,1\}) \Longrightarrow \mu_b = \mu_*.$$

Then,

$$\mu_b = \mu_* \iff \mu_L(b_i) = 0, \ (i \in \{0,1\}).$$

Concerning the left-hand side of the inequality, we have:

$$\begin{cases} S_0^{te} = (S_0^{te} \cap S_1^m) \cup (S_0^{te} \setminus S_1^m) \\ S_1^{te} = (S_1^{te} \cap S_0^m) \cup (S_1^{te} \setminus S_0^m) \end{cases}.$$

In particular, the intersection of the two members for each equation is empty. Then, we can rewrite μ_b as follows:

$$\begin{aligned} \mu_b = &\int_{S_0^{te} \cap S_1^m} \pi_0\rho_0^{te} dg + \int_{S_0^{te} \setminus S_1^m} \pi_0\rho_0^{te} dg + \int_{S_1^{te} \cap S_0^m} \pi_1\rho_1^{te} dg + \int_{S_1^{te} \setminus S_0^m} \pi_1\rho_1^{te} dg \\ &+ \int_{S_0^m \cap S_1^{te}} (\pi_0\rho_0^{te} - \pi_1\rho_1^{te})dg + \int_{S_1^m \cap S_0^{te}} (\pi_1\rho_1^{te} - \pi_0\rho_0^{te})dg. \end{aligned}$$

$$\iff$$

$$\mu_b = \int_{S_1^m \cap S_0^{te}} \pi_1\rho_1^{te} dg + \int_{S_0^m \cap S_1^{te}} \pi_0\rho_0^{te} dg + \int_{S_0^{te} \setminus S_1^m} \pi_0\rho_0^{te} dg + \int_{S_1^{te} \setminus S_0^m} \pi_1\rho_1^{te} dg.$$

As each integrand is positive or null, we have $\mu_b \geq 0$.

1. Let assume that $S_i^m = S_j^{te}$ and $\rho_j^{te} = \rho_j^{te}\mathbf{1}_{\{S_j^{te}\}}$, for $i,j \in \{0,1\}$, $i \neq j$. Then, we have $\mu_b = 0$.

2. Let assume that $\mu_b = 0$. By definition of the different sets, it is easy to show that μ_b is defined as a sum of integrals over disjoint sets. As each integrand is positive or null, it follows that each integral has to be equal to 0. Recalling that $\pi_0\rho_0^{te} > \pi_1\rho_1^{te} \geq 0$ over S_0^{te}, it is obvious that we necessarily have $S_0^{te} \subseteq S_1^m$. For the same reason, we have $S_1^{te} \subseteq S_0^m$ (from the fourth integral). Let $x \in S_1^m \setminus S_0^{te}$. Then, $x \in S_1^m \cap S_1^{te}$ which is impossible because $S_1^{te} \subseteq S_0^m$. It follows that:

$$S_i^{te} = S_j^m, i,j \in \{0,1\}, i \neq j.$$

Then, to ensure that the first two integrals are equal to 0, we necessarily have:

$$\rho_i^{te} = \rho_i^{te}\mathbf{1}_{\{S_i^{te}\}}, i \in \{0,1\}.$$

Finally,

$$\mu_b = 0 \Longleftrightarrow \begin{cases} S_0^m = S_1^{te} \Longleftrightarrow S_1^m = S_0^{te} \\ \rho_i^{te} = \rho_i^{te}\mathbf{1}_{\{S_i^{te}\}} \end{cases}.$$

In other words, the worst case for μ_b is obtained when the model is as bad as possible. ■

Proof of Lemma 3.

$$S_1^{te} \setminus S_1^v = S_1^{te} \setminus (\Omega \setminus S_0^v).$$

Using some set theory properties,

$$S_1^{te} \setminus S_1^v = (S_0^v \cap S_1^{te}) \cup (S_1^{te} \setminus \Omega) = S_0^v \cap S_1^{te} = S_0^v \cap (\Omega \setminus S_0^{te}) = \Omega \cap (S_0^v \setminus S_0^{te}).$$

Then we finally obtain:

$$S_1^{te} \setminus S_1^v = S_0^v \setminus S_0^{te}.$$

■

Proof of Proposition 9.

$$\mu_{te}^c = \int_{S_0^v} \pi_0\rho_0^{te}dg + \int_{S_1^v} \pi_1\rho_1^{te}dg.$$

As $\rho_j^{te} = \rho_j^v$, we have $S_j^{te} = S_j^v$. Then,

$$\mu_{te}^c = \int_{S_0^{te}} \pi_0\rho_0^{te}dg + \int_{S_1^{te}} \pi_1\rho_1^{te}dg.$$

In the incomplete validation case, we have either:

$$\begin{cases} S_1^v \subseteq S_1^{te} \text{ and } S_0^{te} \subseteq S_0^v \\ S_0^v \subseteq S_0^{te} \text{ and } S_1^{te} \subseteq S_1^v \end{cases}.$$

.

By symmetry of the problem, let assume that:

$$S_1^v \subseteq S_1^{te} \text{ and } S_0^{te} \subseteq S_0^v.$$

We then have:

$$\begin{aligned}\mu_{te}^p &= \int_{S_0^v} \pi_0 \rho_0^{te} dg + \int_{S_1^v} \pi_1 \rho_1^{te} dg \\ &= \int_{S_0^{te}} \pi_0 \rho_0^{te} dg + \int_{S_1^{te}} \pi_1 \rho_1^{te} dg + \int_{S_0^v \setminus S_0^{te}} \pi_0 \rho_0^{te} dg - \int_{S_1^{te} \setminus S_1^v} \pi_1 \rho_1^{te} dg.\end{aligned}$$

Using Lemme 3, we have:

$$\mu_{te}^p = \mu_{te}^c - \int_{S_1^{te} \setminus S_1^v} (\pi_1 \rho_1^{te} - \pi_0 \rho_0^{te}) dg.$$

Moreover, we know that $\pi_1 \rho_1^{te} \geq \pi_0 \rho_0^{te}$ over S_1^{te}. Hence, the second term of the previous equation is positive. Then,

$$\mu_{te}^p \leq \mu_{te}^c.$$

■

6.6.1 MV: scores in the incomplete validation set scenario

For this study we make the following assumptions:

- AP behaviour under sodium blockade does not depend on potassium and calcium channel activities.
- AP behaviour under potassium and/or calcium channel blockade are dependent.

The following study is coarse, but presented to justify scores obtained in Section 6.4.2.4 of the manuscript.

Sodium channel blockade

In the incomplete validation case, sodium activities for the validation set belong to $(0.85, 1)$ except *knn* of them, belonging to the other class (i.e. 0.6,0.8). We recall that each activity is an independent realisation of a random variable following a uniform law over (0.6,1). It is then acceptable to assume that all the elements of the test set belonging to (0.85,1) will be well classified (i.e. 37,5% of the segment (0.6,1)). The average, elements of the validation set belonging to the other class (i.e. (0.6,0.8)) is 0.7. Then, it is reasonnable to assume that all the elements of the test with an activity belonging to (0.6,0.7) will be in average well classified (which corresponds to 25% of (0.6,1)). Finally, on (0.7,0.85) we can assume that as we do not have enough information, and then, we well classify half of

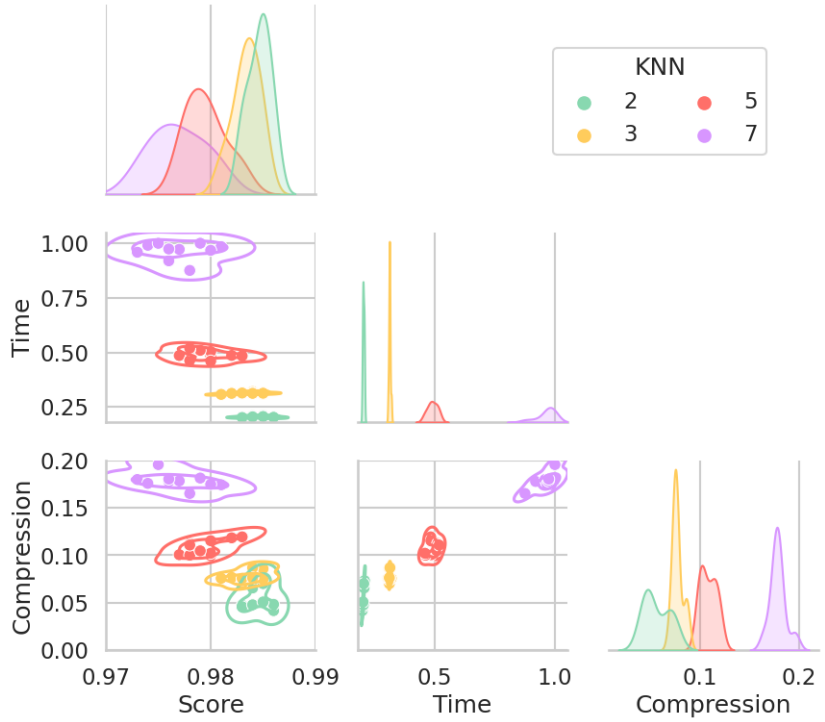

Figure 6.13: Influence of the number of neighbours on the score, the computation time and the compression for the study case 1.

the elements of the test set. In particular, because of the non-linearities of the model it is not possible to assume that the average class delimitation is at $0.5 * (0.7 + 0.85) = 0.775$.

Under the above assumptions, we have the following score:

$$\mu = \frac{0.25 * n_t + \frac{1}{2} * 0.375 * n_t + 0.375 * n_t}{n_t} = 0.8125,$$

where n_t is the number of elements of the test set. Then, by simulation, we expect to have a score close to 0.81 for the sodium channel blockade study in the incomplete validation set case.

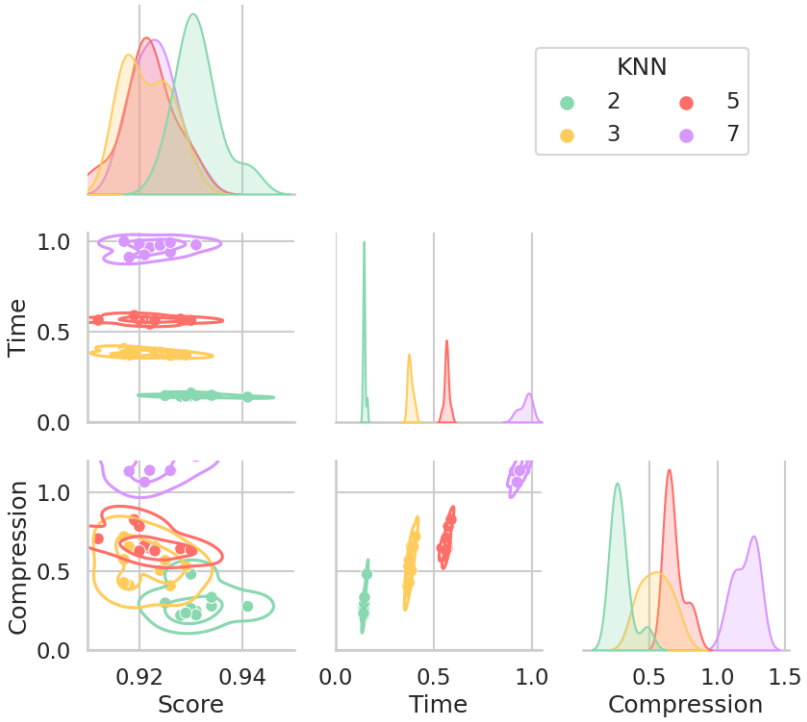

Figure 6.14: Influence of the number of neighbours on the score, the computation time and the compression for the study case 2.

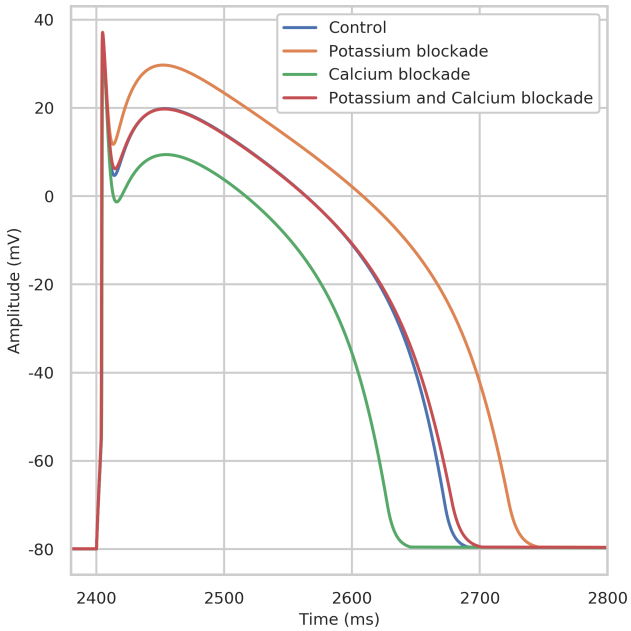

Figure 6.15: Comparison of different channels blockade (20% of blockade). Sodium channel blockade is mainly known to reduce the depolarisation peak, calcium channel blockade is mainly known to reduce the plateau phase and the duration whereas potassium channel blockade is mainly known to induce a signal prolongation.

Chapter 7
Conclusions

This methodological part is devoted to the construction of a machine learning tool with a low number of parameters to tune and easily pluggable into a supervised classification context.

The first section describes a goal oriented double greedy dimension reduction (DGDR) method based on Stiefel manifold properties and the maximisation of a score function related to the classification success rate. The interest of the proposed method is twofold. On one hand, the dimension reduction allows to reduce the overfitting whereas, on the other hand, the success rate on the validation set is maximised. The k_{nn} parameter is the only parameter appearing in the method during the discretisation process.

The second section (based on the same assumptions as in the DGDR method) aims at constructing an augmented set by removing irrelevant or biased samples. This selection is operated in an iterative way by reducing Hausdorff distances between sets. It has been shown that this minimisation is equivalent to the maximisation of the same score function as described in the DGDR method. Moreover, the only free parameter of the method is the same as the one appearing in the DGDR method (i.e. k_{nn} parameter).

For the above reasons, the two proposed strategies seem to be suitable to address the arose problem in the introduction of the manuscript.

Further works could be done to extend and/or improve these proposed methods:

- DGDR technique:
  - Extension to regression tasks (see Section 9).
  - Extension to more than 2 classes.
  - Activation functions can be embedded.
- ASE-HD technique:
  - Cross-validation techniques could be considered in order to improve the robustness of the augmented set construction.
  - Hausdorff distance consideration during the pruning phase to remove the random process.

Despite these improvement opportunities, the proposed methods have shown encouraging results in several study cases:

- DGDR technique: better results (in average) than usual classification and reduction techniques.
- ASE-HD technique: relevant augmented training set with a low sample size.

The following chapters are a direct application of the proposed methods in a safety pharmacological context.

Part III

Patch-clamp studies

Chapter 8
Introduction

Cardiac electrophysiology is a wide field of study at different physical scales: from cell level to body level. In the context of safety pharmacology, many *in vitro* studies are performed at early stages during drug development. These steps are crucial to pursue development phases at higher scales before being released on the market.

These in vitro techniques are essentially patch-clamp techniques (see Section 2.3.1). They are nowadays the gold standard to study compound effects on cardiomyocytes (or induced pluripotent stem cells derived cardiomyocytes). These techniques consist in recording ionic currents passing from either side of membrane cells (sarcolemma in our case of cardiomyocytes studies).

In this context, two collaborations were performed:

- University of Zaragoza, Spain: Esther Pueyo, David Adolfo Sampedro-Puente, Jesus Fernandez-Bes & Pablo Laguna: This work aims at estimating ionic channel activities from Action Potential (AP) signals at control case and under $\beta-$adrenergique stimulation. This work led to the publication of a paper in IEEE Journal of Biomedical and Health Informatics [SPRFB+20].
- Sophion Bioscience, Ballerup, Denmark: Kadla Røskva Rosholm, Lasse Homann & Anders Lindqvist: The goal of this study is to combine automated patch-clamp techniques on multiwell plates to a Hit/No Hit classification. Roughly speaking, the plate allowing to perform multiple patch clamp recordings at the same time permits a high throughput screening. It is then crucial to efficiently and quickly classify whether a compound at a certain concentration has a significant impact on the signal with respect to the control case (i.e. without compound addition).

Chapter 9

Channel activity estimation

Objective: Elevated spatio-temporal variability of human ventricular repolarisation has been related to an increased risk for ventricular arrhythmias and sudden cardiac death, particularly under $\beta-$adrenergic stimulation ($\beta-$AS). This work presents a methodology for theoretical characterisation of temporal and spatial repolarisation variability at baseline conditions and in response to $\beta-$AS. For any measured voltage trace, the proposed methodology estimates the parameters and state variables of an underlying human ventricular action potential (AP) model by combining the DGDR method and the Unscented Kalman Filter (UKF). Such theoretical characterisation can facilitate subsequent characterisation of underlying variability mechanisms.

Material and Methods: *In silico* AP traces were generated based on the ORd action potential model. The ionic conductances of a given AP trace were estimated by the DGDR method extended to regression problems. Those estimates served to initialise and update the ionic conductance estimation, from the UKF method, which is based on the formulation of an associated non-linear state-space representation and the joint estimation of model parameters and state variables. Similarly, $\beta-$AS-induced phosphorylation levels of cellular substrates were estimated by the DGDR-UKF methodology. Performance was tested by building an experimentally calibrated population of virtual cells, from which synthetic AP traces were generated for baseline and $\beta-$AS conditions.

Results: The combined DGDR-UKF methodology led to 25% reduction in the error associated with estimation of ionic current conductances at baseline conditions and phosphorylation levels under $\beta-$AS with respect to individual DGDR and UKF methods. This improvement was not at the expense of higher computational load, which was diminished by 90% with respect to the individual UKF method. Both temporal and spatial AP variability of repolarisation were accurately characterised by the DGDR-UKF methodology.

Conclusions: A combined DGDR-UKF methodology is proposed for parameter and state variable estimation of human ventricular cell models from available AP traces at baseline and under $\beta-$AS. This methodology improves the estimation performance and reduces the convergence time with respect to individual DGDR and UKF methods and renders a suitable approach for computational characterisation of spatio-temporal repolarisation variability to be used for ascertainment of variability mechanisms and its relation to arrhythmogenesis.

Contents

9.1 Introduction

Clinical, experimental and computational studies have demonstrated the important role of cardiac spatio-temporal variability in electrical function, whether it is at cellular level through AP signals or at body scale through electrocardiograms. On one hand, spatial variability refers to electrophysiological differences between cardiac cells or regions of cells and has been to some extent attributed to distinct ionic current contributions to individual APs [PCV+11, SKH+15, LKMGG07, LFYC98, SBOW+14]. On the other hand, temporal variability refers to AP differences between cardiac beats and has been suggested to arise from random fluctuations in ionic currents as well as variations in intracellular calcium handling [PCV+11, LKMGG07, AJD+15, KSH+15, LdLK11, TGOW05]. Numerous investigations have associated elevated spatial and/or temporal variability of ventricular repoloraziations with pro-arrhythmicity and sudden cardiac death [HBT+10, VWvdH+12, SWY+11, WSH+15, SPFBS+20]. In particular, $\beta-$adrenergic stimulation ($\beta-$AS) has been shown to produce exaggerated increases in beat-to-beat variability of repolarisation (BVR), particularly under conditions of reduced repolarisation reserve [JHB+13, JHP+10, GVdWvdL+07]. *In vitro* experiments in isolated cardiomyocytes have suggested that this elevation in BVR by $\beta-$AS is a relevant contributor to arrhythmogenesis by the development of afterdepolarisations and triggered activity [JHB+13, SKH+15, JHP+10, HBI+18, HZJ+13]. In an *in vivo* animal model of long QT1 syndrome, $\beta-$AS has been shown to induce increments in both temporal and spatial dispersion of repolarisation and to facilitate the development of early afterdepolarisations (EADs) and left ventricular aftercontractions, altogether providing the substrate and triggers for the ignition of Torsade de Pointes, a life-threatening ventricular arrhythmia [GVdWvdL+07]. Computational investigations have further contributed to shed light into the mechanisms underlying the relationship between $\beta-$AS induced elevation in BVR and pro-arrhythmic risk [LKMGG07, LdLK11, HZJ+13, PORT16, SPFBP+19]. Nevertheless, most of the computational approaches employed so far in the literature do not account for realistic modelling of cell-to-cell or beat-to-beat AP differences, which should be fundamental to better understand the relationship between BVR and arrhythmogenesis and its modulation by $\beta-$AS.

For the above reasons, the development of stochastic cardiac computational AP models fed with information acquired from human cells or tissues becomes of major interest. In recent years, different methodologies have been proposed to integrate information from cardiac AP signals, or from a set of markers derived from them, by identifying the values of parameters and/or state variables of an underlying electrophysiological model. This allows obtaining a population of virtual AP models representative of a set of experimental data of interest, with the advantage of facilitating assessment of the causes and consequences of BVR by simultaneous assessment of voltage and ionic currents/concentrations. In [LDC+18, TLRG17, CSC15], methodologies based on Genetic Algorithms, Moment Matching and Gaussian Process Emulators were designed for parameter identification at a population level, thus allowing to reproduce the overall AP characteristics in the investigated cell population but hampering individual identification

of the parameters associated with each cellular AP trace. In [JCB+16, JBG+16] ionic parameters were estimated from voltage signals by using Markov Chain Monte Carlo (MCMC)-based methods, which enable parameter estimation for each individual cell. However, on top of the high computational load associated with these methods, they do not account for beat-to-beat variability and do not provide an estimation for other non-measurable state variables of the model, such as ionic concentrations or channel open probabilities, as neither do the methods proposed in [LDC+18, TLRG17, CSC15]. In [SPFBV+19], a methodology based on non-linear state-space representations [Sär13] and the Unscented Kalman Filter (UKF) [JU04] was proposed to identify the parameters and state variables of stochastic human ventricular AP models. This methodology provided robust one-to-one model parameter and state estimation for each AP trace individually, but the computational load was high and it required a long AP signal for accurate estimation.

On the basis of the above described limitations, a methodology for AP model parameter and state estimation that combines fast methodologies based on biomarkers' information with other more complex methodologies based on AP traces' information could be most useful. When condensing AP data into a set of biomarkers, it is important to keep a sufficient amount of information to avoid any risk of degradation in the estimation. To ensure this, the number of computed quantities per sample to build the dictionary, can be potentially substantial (hundreds, thousands, ...) and even higher than the sample size.

In this regime, various phenomena can appear, referred to as the curse of dimensionality [Bel15], which requires data processing to improve classification or regression. The potency of the DGDR method to reduce the dimension in a goal-oriented way is extended to solve regression tasks and then perform parameters estimation. It follows that DGDR method automatically builds a parameter-dimensional regressor input through a sparse linear combination of the dictionary entries to prevent any overfitting risk [Haw04].

The present *in silico* study proposes the combined use of DGDR and UKF-based methodologies to extract information from AP signals at baseline and under $\beta-$AS. Initial DGDR parameter estimates are used to initialise and/or update subsequent UKF estimates so as to facilitate that these remain close to their actual values. To assess the performance of the proposed methodology, a population of stochastic ventricular human cell models is constructed and used to run simulations at baseline conditions and following $\beta-$AS. Methodological performance is first tested over the synthetic AP signals generated for baseline conditions, from which a set of ionic current conductances are inferred for each virtual cell. In a second step, the methodology is tested over synthetic AP signals of the same population following $\beta-$AS, from which the phosphorylation levels of a set of cellular substrates are inferred, considering the previously identified ionic conductances.

This work thus provides a tool to assess mechanisms underlying cardiac spatio-temporal variability and identify those with pro-arrhythmic potential.

9.2 Methods

9.2.1 Stochastic AP Models at Baseline and under $\beta-$AS

9.2.1.1 Stochastic Human Ventricular ORd Model

A stochastic version of the OHara (ORd) human ventricular epicardial AP model [OVVR11] was developed to reproduce experimentally observed BVR. Following the subunit-based approach described in [PCV+11], the set of ordinary differential equations (ODEs) describing ion channel gating for the four principal currents active during AP repolarisation, namely I_{Ks} (slow delayed rectifier potassium current), I_{Kr} (rapid delayed rectifier potassium current), I_{to} (transient outward potassium current) and I_{CaL} (L-type calcium current), were transformed into stochastic differential equations (SDEs) by adding a stochastic term of the form shown in Equation (9.1) for a generic ionic gate x_g, where x_{g_∞} is the steady-state value and τ_{x_g} is the time constant.

$$dx_g = \frac{x_{g_\infty} - x_g}{\tau_{x_g}} dt + \frac{\sqrt{x_{g_\infty} + (1 - 2x_{g_\infty})x_g}}{\sqrt{\tau_{x_g} N_g}} dw. \tag{9.1}$$

The added stochastic term containing the increments of a Wiener process (dw) multiplied by a factor inversely proportional to the number of ion channels (N_g) of the corresponding type was added to the deterministic term defining x_g gating. By including this stochastic term with an accurately estimated number of channels, realistic fluctuations in the ionic gates and the whole-cell ionic currents are reproduced (as a Brownian motion), which are the source for BVR in cellular AP. The number of channels N_g associated with I_{Ks}, I_{Kr}, I_{to} and I_{CaL} were calculated by dividing the default ionic conductance values in the ORd model by the corresponding single channel conductances reported in literature, as described in [SPFBV+19].

9.2.1.2 $\beta-$Adrenergic Signaling model

$\beta-$AS effects were modelled following the approach described in [PORT16], where a modified version of the Xie et al. model [XGP+13], with definition of graded and dynamic phosphorylation levels of cellular protein kinase A (PKA) substrates, was used. The Xie et al. model was updated from the original $\beta-$adrenergic signalling formulation proposed in [SS10] to slow down the I_{Ks} phosphorylation and dephosphorylation rate constants to fit experimental observations. PKA-mediated phosphorylation of phospholemman was accounted for in [XGP+13] by increasing the $Na^+ - K^+$ (NaK) pump affinity for intracellular Na^+ concentration.

9.2.2 Synthetic Data

A population of stochastic AP models was constructed to reproduce the experimentally reported inter-individual variability in human ventricular electrophysiological properties (a sample is then a sequence of AP signals). An initial population of *in silico* cells was generated by using a Monte-Carlo method in which the conductances of eight main ionic

conductances were varied in the range ±100% of their nominal values in the ORd model, with those currents being: I_{Ks}, I_{Kr}, I_{to}, I_{CaL}, inward rectifier potassium current I_{K1}, sodium current I_{Na}, sodium-calcium exchanger current I_{NaCa} and sodium-potassium pump current I_{NaK}. This corresponded to definition of eight multiplying conductance factors, namely θ_{Ks}, θ_{Kr}, θ_{to}, θ_{CaL}, θ_{K1}, θ_{Na}, θ_{NaCa} and θ_{NaK}, varying between 0 (full blockade) and 2 (activation at twice the baseline activity). These multiplying conductances factors can be seen as the factor of the conductance-block model described in Equation (3.5) of Section 3.2.2 to simulate compound effects on the electrical signal. From the 8000 initially generated samples, only 2373 presenting electrophysiological properties within physiologically limits were retained, with those limits shown in Table 9.1 as determined based on [OVVR11, GLL+11, BBOV+17, GPB10, JVS+08, PMPI+02, SHH+98]. The quantified properties included AP duration (APD) at 90% (APD90) and 50% repolarisation (APD50), resting membrane potential (RMP), peak membrane potential (Vpeak), percentage of change in APD90 after blocking individual ionic currents (ΔAPD90) as well maximal concentrations of intracellular sodium (Na_i^+) and calcium (Ca_i^{2+}). The retained models represent *in silico* cells with distinct ionic properties.

AP characteristic	Min. accept. value	Max. accept. value
Under baseline conditions		
APD_{90} (ms)	178.1	442.7
APD_{50} (ms)	106.6	349.4
RMP (mV)	−94.4	−78.5
V_{peak} (mV)	7.3	–
Under 90% I_{Ks} block		
ΔAPD_{90} (%)	−54.4	62
Under 70% I_{Kr} block		
ΔAPD_{90} (%)	34.25	91.94
Under 50% I_{K1} block		
ΔAPD_{90} (%)	−5.26	14.86
Na_i^+ concentration in baseline conditions		
Max. Conc. (μM)	–	39.27
Ca_i^{2+} concentration in baseline conditions		
Max. Systolic (μM)	–	2.23
Max. Diastolic (μM)	–	0.40

Table 9.1: Calibration criteria applied onto ventricular human cell models.

To simulate a range of potentially different $\beta-$AS effects in the constructed population of stochastic AP models, multiplying factors θ_{fCaL}, θ_{fKs} and θ_{fNaK} for the PKA phosphorylation levels f_{CaL} , f_{Ks} and f_{NaK} were varied so that these phosphorylation levels ranged between the values at baseline (i.e. without Isoproterenol (ISO)) and the values after the application of an ISO dose of $1\mu M$ associated with maximal effects. This population of phosphorylation levels, generated by using a Monte-Carlo method, was

combined with the above described population of stochastic AP models to obtain a global population of size 2373 with 11 simultaneously varying parameters. This population was divided into training and validation subpopulations of sizes 2000 and 373 respectively.

AP traces of 1100 beats were simulated at baseline and following $\beta-$AS, respectively, by applying $1ms$ rectangular stimulus pulses of $52pA/pF$ amplitude delivered at $1Hz$ pacing frequency. The Euler-Maruyama scheme [Mao15] was used to solve the SDEs with an integration time step of $dt = 0.02ms$ that ensured numerical convergence. The last 100 beats of each condition (baseline, $\beta-$AS) were used for further analysis to ensure convergence had been reached.

Independent standard Gaussian noise was added to the synthetically generated AP data, as described in [SPFBV+19], to simulate recording noise as in experimentally acquired data. These noisy APs were input to the estimation methodologies tested in this study.

9.2.3 State-Space Formulation and Augmented States

9.2.3.1 State-Space Formulation

The stochastic version of the ORd model with unknown ionic conductance factors (for baseline conditions) or phosphorylation levels (for $\beta-$AS conditions) was formulated as a non-linear discrete-time state-space model [Sär13] following the approach described in [SPFBV+19]. In these state-space models the only measured variable was considered to be the transmembrane voltage (AP), while there were a number of hidden variables, including ionic concentrations and opening probabilities of ionic gates.

For baseline conditions, model parameters to be estimated were the factors multiplying the nominal conductances of I_{Ks}, I_{Kr}, I_{to}, I_{CaL}, I_{K1}, I_{Na}, I_{NaCa} and I_{NaK}. Hence, the vector of static model parameters was $\theta = \{\theta_{Ks}, \theta_{Kr}, \theta_{to}, \theta_{CaL}, \theta_{K1}, \theta_{Na}, \theta_{NaCa}, \theta_{NaK}\}$, representing variations in the ionic conductances relative to the default values in the ORd model, $g_s = \theta_s g_{s,ORd}$, where $s \in \{Ks, Kr, to, CaL, K1, Na, NaCa, NaK\}$ is the channel species. Note that the same factor θ_s applies to the number of ion channels of each species: $N_s = \theta_s N_{s,ORd}$, as the unitary conductance of each ionic species was assumed to be constant, based on reported experimental findings [FKF+92].

For $\beta-$AS conditions, model parameters to be estimated were the factors multiplying the phosphorylation levels of the PKA substrates whose phosphorylation had a remarkably higher impact on the AP, which were I_{Ks}, I_{CaL} and I_{NaK} currents [NML+15], in agreement with findings reported for other $\beta-$adrenergic signalling models [HVWR11]. Consequently, the vector of static model parameters was $\theta = \{\theta_{fKs}, \theta_{fCaL}, \theta_{fNaK}\}$, representing variations in the phosphorylation levels f_{K_s}, f_{CaL} and f_{NaK} relative to the default values in the modified Xie model, $f_s = \theta_s f_{s,Xie}$, where $s \in \{K_s, CaL, NaK\}$. For both baseline and $\beta-$AS conditions, the vector θ of model parameters was estimated for each given input AP trace.

The state-space representations used in this study were of the form:

$$x(k) = f\Big(x(k-1), q(k-1), \theta\Big), \tag{9.2}$$

$$y(k) = h\Big(x(k)\Big) + r(k), \tag{9.3}$$

where the process equation (see Equation (9.2)) was defined by a non-linear function $f(\cdot)$ with three different input vectors: $x(k)$, containing the state variables of the stochastic AP model; $q(k)$ representing non-additive process noises related to Wiener increments; and θ containing the model parameters to be estimated. On the other hand, the measurement equation (see Equation (9.3)) was defined by the function $h(\cdot)$ relating the measured variable (transmembrane voltage) with the vector of the model state variables. In this study, $y(k) = v(k) + r(k)$, where $v(k)$ represents the noiseless AP and $r(k)$ was assumed to be an additive white Gaussian noise.

9.2.3.2 Augmented State-Space

To perform joint estimation of model parameters and state variables for a given input noisy AP, the state-space representation of Equation (9.2)-(9.3) was reformulated as described in [SPFBV+19]. In brief, *state augmentation* [Sär13] was applied to convert the static parameter vector θ into a time-varying parameter vector $\widetilde{\theta}(k)$ using a random walk model with drift:

$$\widetilde{\theta}(k) = \widetilde{\theta}(k-1) + \delta(k),$$

where $\delta(k)$ represents an artificial noise whose components, defined by i.i.d. zero-mean Gaussian processes with very small variance. An augmented state vector $z(k)$ was built joining the state variable vector $x(k)$ with the new parameter vector $\widetilde{\theta}(k)$ and the process noise vector $q(k)$:

$$z(k) = \Big[x(k), q(k), \widetilde{\theta}(k)\Big]^T.$$

The previous process (see Equation (9.2)) and measurement equations (see Equation (9.3)) were replaced with:

$$z(k) = f_a(z(k-1)) + \epsilon(k), \tag{9.4}$$

$$y(k) = h_a(z(k)) + r(k), \tag{9.5}$$

where f_a and h_a are the augmented versions of f and h respectively, and $\epsilon(k)$ contains noises related to the Wiener increments of the stochastic AP model represented by $q(k)$ and to the new parameter vector $\widetilde{\theta}(k)$ represented by $\delta(k)$.

9.2.4 Individual and Combined DGDR- and UKF-based Methods

9.2.4.1 DGDR and Dictionary entry computations

Dictionary entry computations: For this study the dictionary entry size is $n_g = 889$. For each beat, the following quantities are computed: APD_{30}, APD_{50}, APD_{90}, LTV,

NLTV, NSTV, STV, Trian, Vpeak, Vrest and dVdtmax where definitions are given in Table 9.2.

Quantity	Definition
APD_X	AP duration at $X\%$ of repolarisation [TLRG17]
LTV	Long term variability [PDB+18, SPFBV+19]
NLTV	N-beats window Long term variability [SPFBP17]
STV	Short term variability [PDB+18, SPFBV+19]
NSTV	N-beats window Short term variability [SPFBP17]
Trian	Triangulation: $APD_{90} - APD_{30}$
Vpeak	Maximum depolarisation
Vrest	Resting potential
dVdtmax	Maximum value of the AP derivative over time

Table 9.2: Definitions of extracted quantities.

The average and standard deviation of these quantities (over the 100 beats) shape the first 22 entries. Then, pairwise products are introduced to consider non-linearities, leading to $n_{g_1} = 22 + 253 = 275$ entries. Then, the same process was performed on wavelet coefficients (i.e. extraction for each beat and average and standard deviation computation over the 100 beats). A total of $n_{g_2} = 614$ entries are related to wavelet coefficients. It follows that the total dictionary entry size is $n_g = n_{g_1} + n_{g_2} = 889$. The same dictionary is computed for control cases (keeping the stochastic process, but with multiplying factors for ionic conductances set to 1). Finally ratios with respect to the control case are considered to build the final dictionary G.

DGDR extension to regression problems: The DGDR method was used to estimate the parameters of the stochastic AP model, which represent part of the components of the augmented state vector $z(k)$. DGDR is based on high-dimensional data analysis and aims at mitigating the curse of dimensionality [Bel15] by projecting data into a low subspace through a sparse linear combination of the dictionary entries. In [LR19], data projection is performed such that a classification success rate is maximal, which can be achieved by maximising a score function based on the distributions of the projected data of each class. To apply the DGDR method to regression problems, the score function was replaced by an ℓ^2 norm that minimises the error between the actual values of the ionic conductances or phosphorylation levels and a sparse linear combination of the dictionary entries in a training set:

$$\omega_* = \arg\min_{\omega_i} || \sum_{i=1}^{n_g} \omega_i G^{(i)} - \theta_c ||_{\ell^2},$$

where ω_i are the weights to be determined, n_g is the number of dictionary entries, $G^{(i)}$ is the i^{th} dictionary entry of the training set and θ_c are the known values of the parameters in the training set.

As in [LR19], the early stopping criterion was applied on a validation set to avoid overfitting risk, which leads to a sparse combination of the dictionary entries and the weight vector ($||\omega||_{\ell^0} << n_g$). Thus, given a new AP trace in the validation set, the learned linear combination was applied to estimate the model parameters. For this study from the $n_g = 889$ extracted entries, 100 were selected for the linear combination ($||\omega||_{\ell^0} = 100$) as this already led to improvements in all estimation errors below 10^{-3} when adding a new dictionary entry over the training set.

The linear combination of 100 entries was a good choice to minimise the cost function in the training set while avoiding overfitting in the validation set. As expected, the dictionary entries selected for the estimation of each model parameter were strongly related to the AP phase where the ionic conductances or phosphorylation factors have a more dominant role. As an example, the most relevant biomarkers for estimation of θ_{Na} and θ_{K1} were related to the AP upstroke velocity and RMP, respectively (see right panel of Figure 9.2).

A learning phase was separately performed for each of the model parameters to be estimated. The selected dictionary entries were not the same, which is a direct consequence of the goal-oriented concept of the DGDR method and ensures a certain explanation of the selected entries. The full process for the training step took around 3 hours on about 50 processors for the estimation of the eight ionic current conductances at baseline and proportionally less for the estimation of the three phosphorylation levels under $\beta-$AS. Once the learning phase was performed, the estimation of a new sample was immediate by scalar product between the sample vector and the weight vector (computed at the learning phase).

This training process was performed over a population of 2000 models while evaluation was carried out over 273 models, leading to adequate levels of accuracy. Figure 9.1 (left panel) illustrates an example of θ_{CaL} estimation by DGDR, showing the uniform dispersion of the point cloud that provides a measure of the uncertainty in the estimation. In addition, the DGDR method led to similar accuracy levels for training and evaluation populations as can be observed in Figure 9.1 (right panel) where the distribution of the absolute error between the actual (θ_{CaL}) and estimated ($\widetilde{\theta}_{CaL}$) parameter values is shown.

Similar results were obtained in the estimation of the other model parameters and are summarised in Figure 9.2.

Once the training phase is performed, estimated densities on the error (see the middle panel of Figure 9.2) are quite similar between the training and the validation set, meaning that the overfitting is weak. The estimation of θ_{Na}, θ_{Kr}, θ_{K1} and θ_{CaL} are much easier to estimate using DGDR than other parameters as the standard deviation error is lower. θ_{Ks} estimation is the worst. Most weighted selected entries are relevant with respect to its estimated parameter (see right panel of Figure 9.2).

These results served to support the adequacy of separating the population into a training set of 2000 models and a validation set of 373 models.

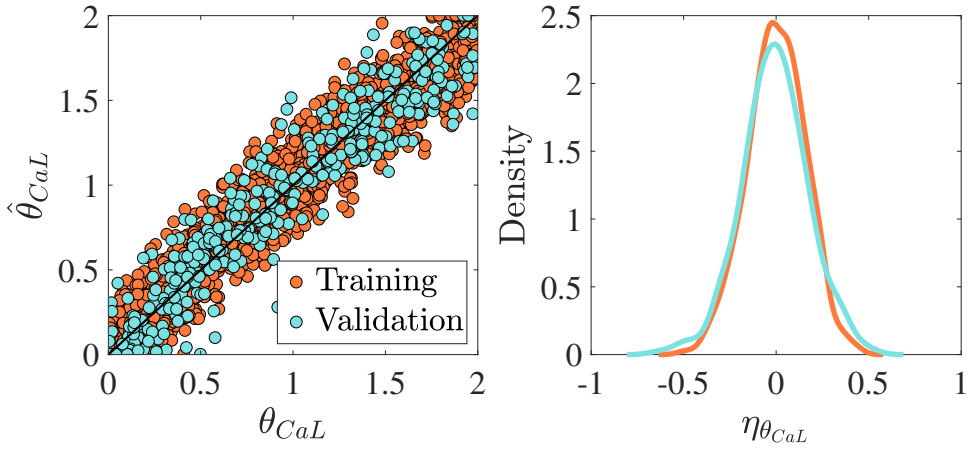

Figure 9.1: Left panel: Estimated ($\hat{\theta}_{CaL}$) vs actual (θ_{CaL}) values of the factor multiplying maximal g_{CaL} in the training and validation populations. Right panel: Density of the absolute error in the estimation of θ_{CaL} for the training and validation sets.

9.2.4.2 UKF

The UKF [JU04] was used to estimate the states of the non-linear state-space formulation described by Equation (9.4)-(9.5), which provides estimates for the parameters and state variables of a stochastic ventricular human cell model for any given AP trace. The values of three UKF setting parameters, commonly denoted by α, β and κ, were set to define the spreading of Sigma-Points around the mean state estimates (controlled by α and κ) and to reflect prior knowledge of states' statistical distributions (controlled by β). In this work, $\alpha = 1$, $\beta = 0$ and $\kappa = 3 - L$ (with respect to [WVDMH01]), where L is the number of states ($L = 71$ for baseline and $L = 68$ for $\beta-$AS conditions). This led to a value for the spread of the state covariance matrix corresponding to $\sqrt{\gamma} = 1.7321$, in accordance with feasible values [SCS12], and to sums of weights of means and covariances equal to one:

$$\begin{cases} \sum_{i=0}^{2L} W_i^{(c)} = 1 \\ \sum_{i=0}^{2L} W_i^{(m)} = 1 \end{cases} .$$

Two additional hyper-parameters were set in the UKF implementation, which determine the process noise variance σ_θ^2 (the same for all components of the model parameter vector) and the measurement noise variance σ_r^2. A range of values for σ_θ^2 were tested and the one rendering best performance was selected. The value for σ_r^2 was set to $1mV$ [SPFBV+19].

The initialisation of the mean and covariance matrix of the state vector was obtained from the training population. The state variables related to stochastic AP model parameters (representing multiplying factors for ionic conductances baseline and for phosphorylation levels under $\beta-$AS) were constrained to remain in the interval $[0, 2]$.

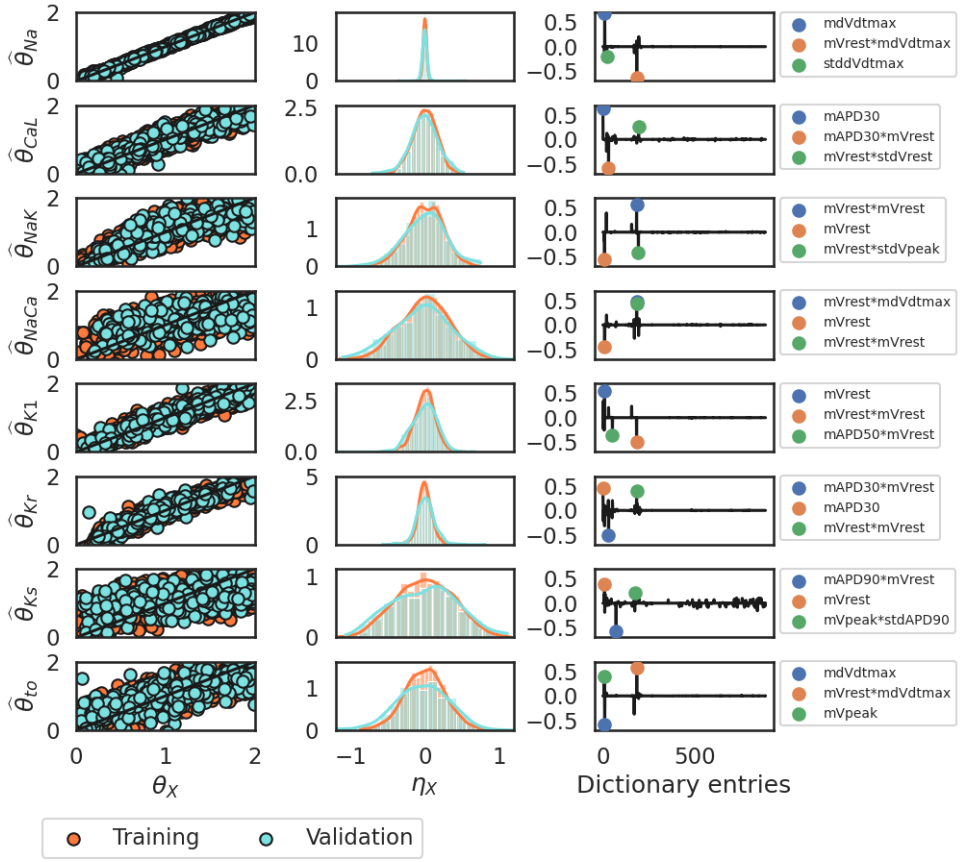

Figure 9.2: Left panel: Estimated ($\hat{\theta}_c$) vs actual corresponding (θ_c) values of the factor multiplying maximal g_c in the training and validation population sets. Middle panel: Density of the absolute error in the estimation of θ_c for the training and validation sets. Right panel: Weights obtained with DGRD with respect to $\widehat{\theta}_c$.

9.2.4.3 Combined UKF-DGDR

DGDR and UKF methods were combined to enhance their individual characteristics in terms of estimation accuracy and computational costs. In particular, DGDR was used for initialisation and updating of UKF estimation to take parameter estimates closer to their actual values and to avoid local minima in the estimation:

- Initialisation (INI): The model parameter estimates obtained by DGDR were used to initialise the corresponding elements of the state vector, which was subsequently estimated by UKF. DGDR provided estimates for both the mean of the parameter vector, $\widehat{\theta}^{DGDR}$, and its covariance matrix, P^{DGDR}.

- Updating (UP): The model parameter estimates obtained by DGDR were used to update the UKF-based parameter estimation in each cardiac cycle. At the end of

each cycle, the corresponding elements of the state vector estimated by UKF (mean $\widehat{z}_k$ and covariance matrix P_k) were updated according to the estimates for the mean $\widehat{\theta}^{DGDR}$ and covariance matrix P^{DGDR} obtained by DGDR as follows:

$$\begin{cases} d = \widehat{z}^{DGDR} - H\widehat{z}_k \\ \mathcal{S} = HP_kH^T + P^{DGDR} \\ \mathcal{K}_{up} = P_kH^TS^{-1} \end{cases},$$

with $\widehat{z}^{DGDR} = \left[O_{N_x}\widehat{\theta}^{DGDR}O_{N_q}\right]$, where O_{N_x} is a $N_x \times 1$ zero vector, O_{N_q} is a $N_q \times 1$ zeros vector and H is a $(N_x + N_\theta + N_q) \times (N_x + N_\theta + N_q)$ matrix of 0 values everywhere, except for the last $N_\theta \times N_\theta$ submatrix occupied by an identity matrix. In the above, N_θ is the number of model parameters, N_x is the number of model state variables and N_q is the number of Wiener processes.

The UKF-based updated estimates for the mean and the covariance matrix of the state vector were:

$$\begin{cases} \widehat{z}_{k_{up}} = \widehat{z}_k^- + \mathcal{K}_{up}d \\ P_{k_{up}} = (I_{N_\theta} - \mathcal{K}_{up}H)P_k \end{cases}.$$

9.2.5 Performance Evaluation

The performance of DGDR, UKF and their combination was evaluated for estimation of eight ionic current conductances at baseline conditions and for estimation of three phosphorylation levels under $\beta-$AS conditions. In the latter case, the values for the eight ionic conductances were set at those estimated at baseline. The estimation performance was evaluated by [SPFBV+19]:

9.2.5.1 AP estimation

The root mean square error between the original noiseless AP trace and the estimated AP trace was calculated over the last 5 cycles (a larger number of cycles did not improve the estimation performance [SPFBV+19]):

$$\xi_v = \sqrt{\frac{1}{K_N}\sum_{k=0}^{K_N-1}|v(k) - \widehat{v}(k)|^2},$$

where K_N is the number of samples contained within the last $N = 5$ cycles.

9.2.5.2 State and parameter estimation

The mean absolute error between the actual and estimated values of each state was calculated over the last 5 cycles:

$$\eta_{z_j} = \frac{1}{K_N} \sum_{k=0}^{K_N-1} |z_j(k) - \widehat{z}_j(k)|,$$

where z_j is the actual value of the state variable j and $\widehat{z}_j$ is the estimated value, with $j = 1, \ldots, L$, being L the length of the augmented state vector $z(k)$.

A global accuracy measurement $\bar{\eta}_\theta$ of model parameter estimation was defined as the average of the mean absolute errors η_{θ_i}, $i = 1, \ldots, N_\theta$, corresponding to all estimated model parameters:

$$\bar{\eta}_\theta = \frac{1}{M} \sum_{\theta' \in \theta} \eta_{\theta'},$$

where $\eta_{\theta'}$ is the mean relative error for model parameter $\theta' \in \theta$ and $M = 8$ (for conductance factors) or $M = 3$ (for phosphorylation factors).

9.3 Results

9.3.1 Implementation of UKF method

The performance of the UKF method as a function of the process noise standard deviation σ_θ is illustrated in Figure 9.3, which shows the mean parameter estimation error in the ORd model when varying σ_θ by several orders of magnitude.

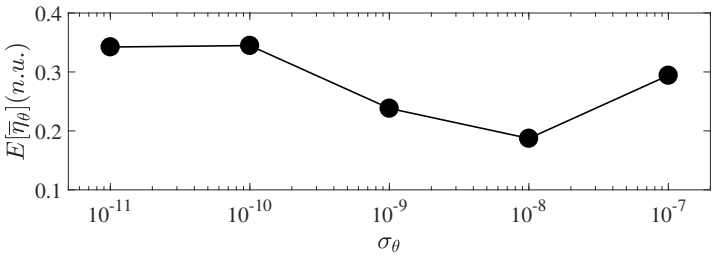

Figure 9.3: Average of mean absolute parameter estimation error $\mathbb{E}[\bar{\eta}_\theta]$ in the ORd model as a function of the standard deviation of the process noise σ_θ.

The minimal average error $\mathbb{E}[\bar{\eta}_\theta]$ was achieved for $\sigma_\theta = 10^{-8}$, which was used for all subsequent analyses. In the case of the root mean square error in AP estimation, ξ_v, its values were minimally affected by the choice of σ_θ for all tested σ_θ values. In the following sections the estimation performances of the DGDR and UKF methods individually and in combination are presented.

9.3.2 Combined DGDR and UKF Methods: Initialisation Effects

The use of the estimates obtained by DGDR for the mean $\widehat{\theta}^{DGDR}$, and the covariance matrix P^{DGDR}, of the model parameter vector as initialisation for the UKF method led

to two important benefits. On one hand, it reduced the time required for the estimates to reach convergence, in turn diminishing the computational cost. On the other hand, it led to more accurate estimates, as shown in Figure 9.4 for the estimation of θ_{Na} in one of the models of the population at baseline conditions.

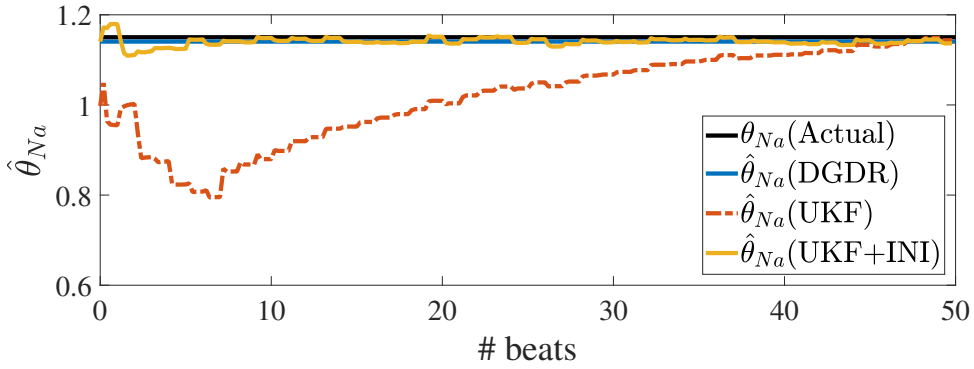

Figure 9.4: Example of actual θ_{Na} value and time course of $\hat{\theta}_{Na}$ as estimated by DGDR, UKF and UKF+INI methods for a virtual cell at baseline.

While for the individual UKF method more than 40 beats were required for the estimation error to be below 0.04, when the combined UKF+INI method was used the number of required beats was 5 for that particular example.

9.3.3 Combined DGDR and UKF Methods: Updating Effects

The use of the estimates obtained by DGDR for the mean $\widehat{\theta}^{DGDR}$, and the covariance matrix, P^{DGDR}, of the model parameter vector to update the UKF estimation at the end of each beat helped to retain the parameter estimates close to the actual parameter values and to reduce the uncertainty in the estimation, as confirmed by a reduction in the estimation covariance matrix. Additionally, this UKF+UP approach diminished the convergence time and, thus, the associated computational cost. The benefit of using the DGDR-derived estimates for UKF updating is illustrated in Figure 9.5 for the estimation of θ_{Kr} in one of the models of the population at baseline conditions.

When only UKF is employed, the parameter estimates may fall in a local minimum and may never reach a value close to the actual one. As can be observed from the figure, the UKF and UKF+UP estimates were the same for the first beat whereas the updating subsequently led to remarkably enhanced results.

9.3.4 Performance Comparison

The performances of the individual DGDR and UKF methods and their combinations, either by initialisation and/or updating, were assessed in terms of the average mean $\mathbb{E}[\bar{\eta}_\theta]$ and standard deviation $\mathbb{E}[\bar{\sigma}_{\eta_\theta}]$ of the absolute error. Top panel of Figure 9.6 illustrates $\mathbb{E}[\bar{\eta}_\theta]$ for the five evaluated methods at baseline conditions.

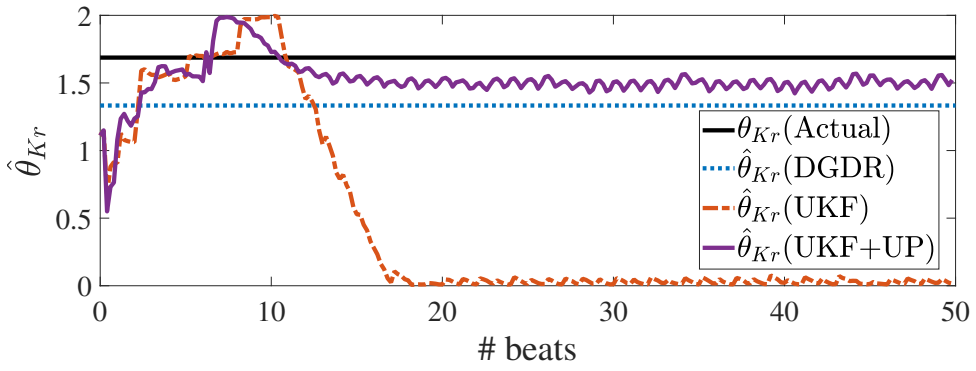

Figure 9.5: Example of actual θ_{K_r} value and time course of $\widehat{\theta}_{K_r}$ as estimated by DGDR, UKF and UKF+UP methods for a virtual cell at baseline.

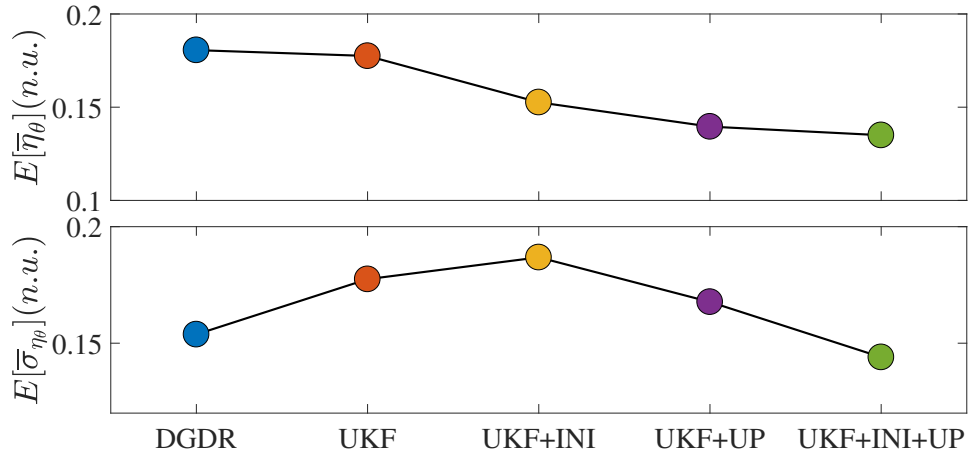

Figure 9.6: Average over the validation population at baseline of mean (top panel) and standard deviation (bottom panel) of absolute parameter estimation error $\bar{\eta}_\theta$ for the five evaluated methods.

As it can be seen from Figure 9.6, the individual DGDR and UKF methods led to approximately the same level of error ($\mathbb{E}[\bar{\sigma}_{\eta_\theta}]$ values of 0.1806 and 0.1775, respectively), with a larger associated computational cost in the case of the UKF method. The combination of DGDR and UKF remarkably improved the estimation performance, either when combined through initialisation or through update and, particularly, when combined through both ($\mathbb{E}[\bar{\sigma}_{\eta_\theta}]$ values of 0.1526 for UKF+INI, 0.1396 for UKF+UP and 0.1350 for UKF+INI+UP). Bottom panel of Figure 9.6 presents the estimation uncertainty for the five evaluated methods. As can be observed, initialisation and updating by DGDR contributed to reduce the parameter estimation uncertainty of the UKF method.

Figure 9.7 shows boxplots for the mean absolute error in the estimation of each ionic conductance factor by each of the five evaluated methods at baseline conditions.

As can be observed from Figure 9.7, the combined UKF+INI+UP method presents

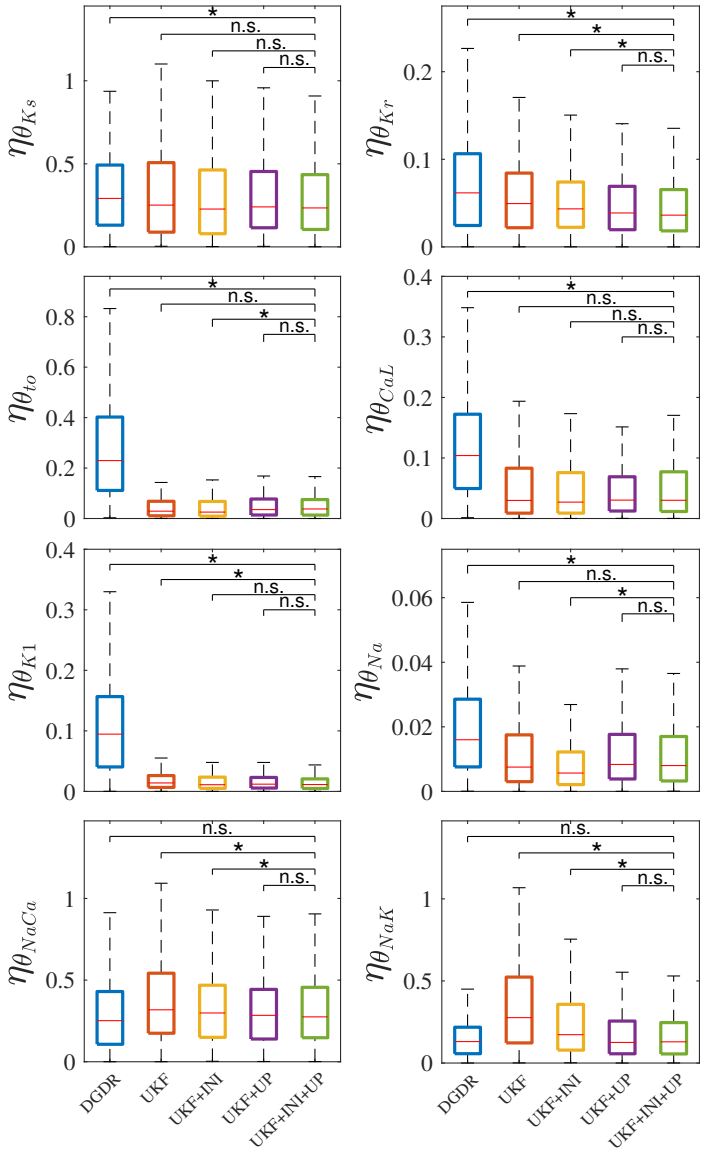

Figure 9.7: Boxplots of absolute estimation errors η_θ for the factors multiplying ionic current conductances calculated for the five evaluated methods. Statistically significant differences by Wilcoxon signed-rank test (p-value < 0.05) are denoted by *, while non-significant differences are denoted by $n.s$, for a number of cells equal to 373.

better performance than the individual DGDR and UKF methods for almost all estimated factors. The most accurate results were obtained for θ_{Na}, with median estimation errors $\widehat{\theta}_{Na}$ being lower than 0.05. On the other hand, the least accurate results were obtained for θ_{Ks}, θ_{NaCa} and θ_{NaK}. Figure 9.8 presents results related to estimation uncertainty.

Figure 9.8, left panel, illustrates the time course of the estimation uncertainty quantified

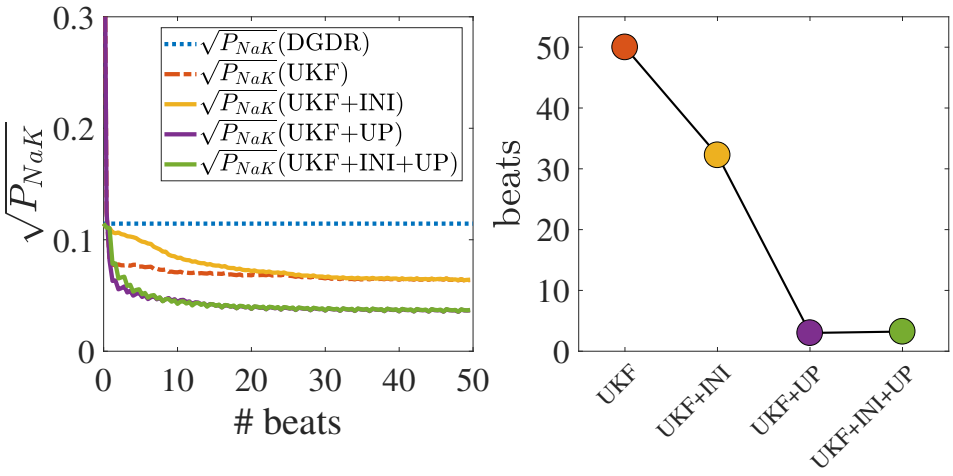

Figure 9.8: Left panel: Time course of estimation uncertainty in terms of square root of covariance matrix $\sqrt{P_{NaK}}$ for each of the five evaluated methods. Right panel: Number of beats required by each evaluated method to reach the same level of accuracy as the UKF method, as quantified by the averaged covariance over all estimated model parameters.

by the square root of the covariance matrix $\sqrt{P_{NaK}}$ in the estimation of θ_{NaK} for one virtual cell in the population at baseline conditions. As can be observed from the figure, the combination of DGDR and UKF presented lower uncertainty than the individual DGDR and UKF methods, with the impact of updating being notably larger than the initialisation. Figure 9.8, right panel, provides an additional characterisation of the estimation uncertainty quantified by the number of beats required by each UKF-based method to reach the same value of the averaged standard deviation of the absolute estimation errors as the individual UKF. The impact of updating on the reduction of the estimation uncertainty as only few beats are needed instead of decades (with the others methods) for a same error.

9.3.5 Replication of AP traces and Biomarkers at Baseline

The performance of the five proposed methods to replicate AP traces at baseline conditions was assessed by generating APs from the ORd model with the different sets of estimated parameters and by comparing them with the input AP traces. Also, the comparison was established in terms of AP-derived biomarkers like APD and STV. Figure 9.9, left panel, shows the probability density function of the differences between the APD from the input AP trace and the APD calculated from the estimated AP trace for DGDR, UKF and UKF+INI+UP.

Similarly, Figure 9.9, right panel, shows results for STV. As can be observed from the figure, the combined UKF+INI+UP method provides the best fitting to the actual data, as confirmed by the fact that the distributions of ΔAPD and ΔSTV are more concentrated around 0. On the other hand, the DGDR method presents reduced accuracy for APD

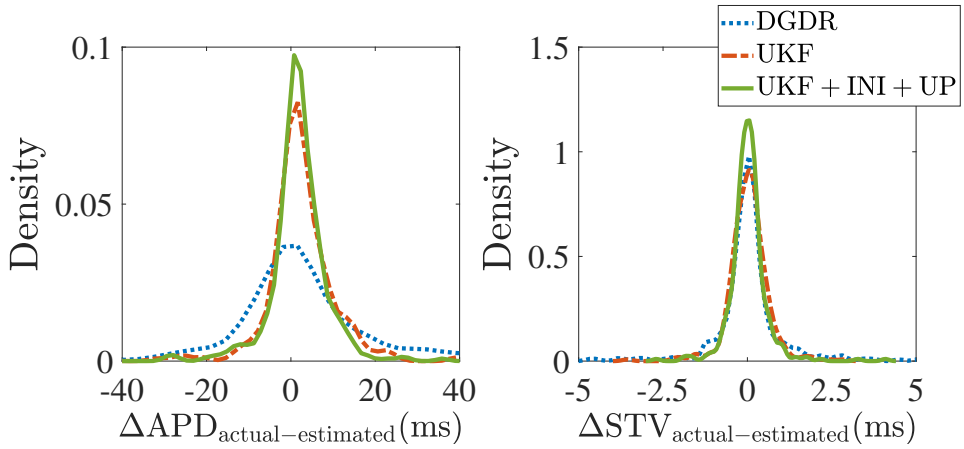

Figure 9.9: Probability density function of ΔAPD (left panel) and ΔSTV (right panel) for the validation population, with ΔAPD (ΔSTV, respectively) calculated as the difference between APD (STV, respectively) from the input AP trace and APD (STV, respectively) from the estimated AP trace for each evaluated method under baseline conditions.

estimation, although very similar to UKF and UKF+INI+UP for STV estimation.

9.3.6 Estimation of Phosphorylation Factors, AP traces and Biomarkers under $\beta-$AS

Considering the ionic conductance estimates obtained for baseline conditions, the next step was to test the performance of DGDR, UKF and UKF+INI+UP (for which the best results were obtained in the previous section) to estimate the phosphorylation levels for the validation population of models under $\beta-$AS effects. Figure 9.10 shows boxplots of the mean absolute errors $\bar{\eta}_\theta$ for the estimation of the three ISO-induced phosphorylation levels.

As can be observed from the figure, the UKF+INI+UP method increased the accuracy in the estimation of θ_{fKs} and θ_{fNaK} with respect to the individual DGDR and UKF methods, whereas for θ_{fCaL} UKF was slightly better in terms of median absolute error, but not in terms of averaged absolute error ($\bar{\eta}_\theta = 0.34$ for both methodologies). Taking together the three estimated factors for the phosphorylation levels and results over the whole validation population, the combined UKF+INI+UP method led to a reduction in the averaged mean absolute error $\mathbb{E}[\bar{\eta}_\theta]$, of 15.29% and 20.01% with respect to the individual use of DGDR and UKF, respectively. The average mean absolute errors, $\mathbb{E}[\bar{\eta}_\theta]$, for ISO-induced phosphorylation level factors were higher (0.38, 0.40 and 0.32 for DGDR, UKF and combination respectively) than those obtained for ionic conductance factors due to the fact that the error in the ionic conductance estimation was propagated into the phosphorylation level estimation.

Figure 9.11, left panel, shows the probability density function of the differences between the APD from the input AP trace and the APD calculated from the estimated AP trace

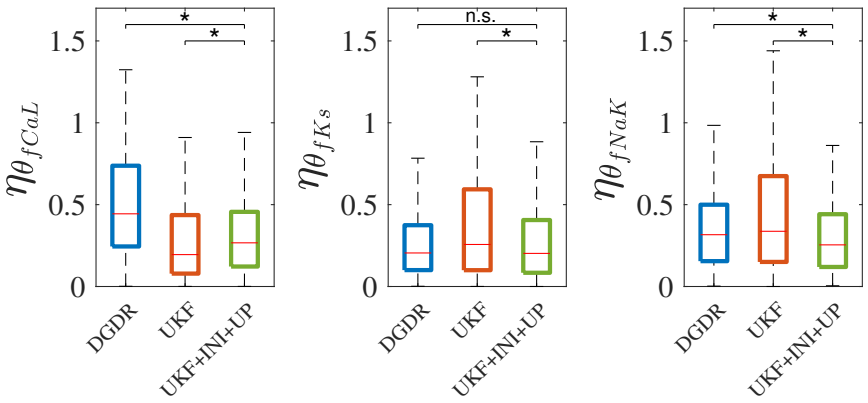

Figure 9.10: Boxplots of absolute estimation errors η_θ for the factors multiplying ISO-induced phosphorylation levels calculated for three evaluated methods. Statistically significant differences by Wilcoxon signed-rank test (p-value < 0.05) are denoted by *, while non-significant differences are denoted by $n.s$, for a number of cells equals to 373.

after estimation of the ionic conductances at baseline and phosphorylation factors under $\beta-$AS for DGDR, UKF and UKF-INI-UP. Figure 9.11, right panel, shows analogous results for STV.

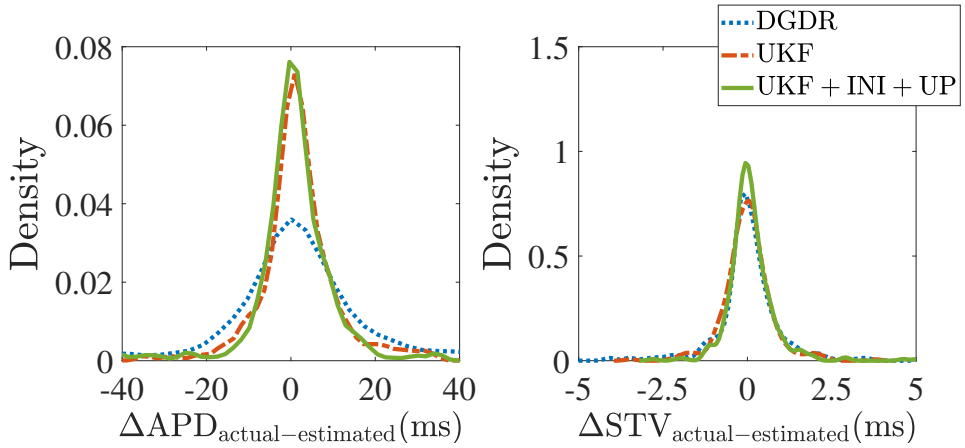

Figure 9.11: Probability density function of ΔAPD (left panel) and ΔSTV (right panel) for the validation population, with ΔAPD (ΔSTV, respectively) calculated as the difference between APD (STV, respectively) from the input AP trace and APD (STV, respectively) from the estimated AP trace for each evaluated method under $\beta-$AS conditions.

Again, the combined UKF-INI-UP provided the best fitting for both ΔAPD and ΔSTV, whereas the DGDR method presented the highest differences between actual and estimated APD and comparable performance to UKF and UKF+INI+UP in the case of STV. As an illustration of the above results, Figure 9.12 shows the actual and estimated APs (mean over 100 beats) calculated from the set of estimated parameters by each of

the evaluated methods for a cell in the validation population.

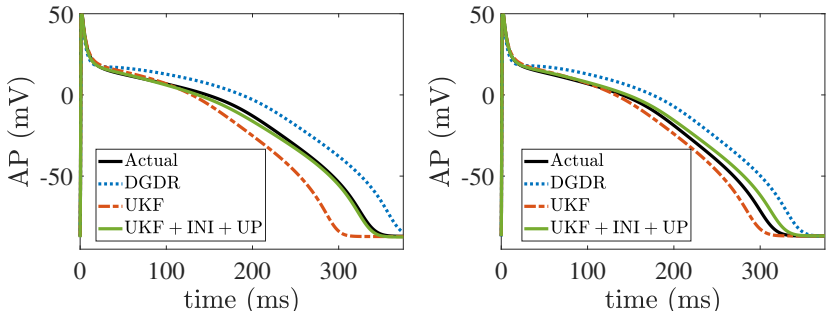

Figure 9.12: Actual and estimated APs (mean over 100 beats) calculated from the set of estimated parameters by each evaluated method at baseline (left panel) and under $\beta-$AS (right panel) for one of the virtual cells in the validation population.

Both at baseline and under $\beta-$AS, the AP estimated by DGDR+UKF remarkably better matched the actual AP as compared to those obtained by DGDR or UKF individually. Not only the mean AP, but also the variability over 100 beats was better reproduced by DGDR+UKF as compared to DGDR and UKF.

9.4 Discussion

A novel approach based on the combined use of the Double Greedy Dimension Reduction (DGDR) and the Unscented Kalman Filter (UKF) has been proposed as a method for joint estimation of parameters and state variables of computational human ventricular stochastic models from given input AP traces. By using this combined methodology, different sets of ionic parameters, namely ionic current conductances and phosphorylation levels of cellular substrates, are estimated for each given individual AP trace at baseline conditions and following $\beta-$AS. The proposed methodology outperforms individual DGDR and UKF methods and has an affordable computational cost. It allows realistic characterisation of spatio-temporal variability at baseline and following $\beta-$AS, thus enabling improved investigation of variability mechanisms and arrhythmic risk prediction. This can prove fundamental to assess the role of $\beta-$AS in leading to exaggerated increases in BVR that facilitate the occurrence of arrhythmic events in certain cases but not in others [SPFBP+19]. In the following, relevant characteristics of the proposed methodology as well as major benefits and shortcomings associated with its use are discussed.

9.4.1 DGDR Method

The DGDR method was used to obtain estimates for the model parameters, which were subsequently fed to the UKF method to build the combined DGDR-UKF method. The intrinsic characteristics of the DGDR method, which include one-to-one matching

between input AP traces and the set of estimated model parameters, ability to manage the stochastic behaviour of the AP traces and low computational burden make this methodology suitable for the problem at hand.

A key factor in the performance of the DGDR method involves a correct training phase. To obtain high levels of estimation accuracy, training should be performed over large populations, which in the case of this study corresponds to a large set of synthetic AP traces. Confirmation on the appropriateness of the training population dimension was provided by the fact that similar estimation errors were attained in both the training and validation populations. If training dimension had not been sufficient, estimation uncertainty in the validation population would have been much greater than that obtained in the training population. The time required to obtain the estimation dictionaries from the training population was just three hours, being subsequent calculation of parameter estimates immediate (scalar product of two vectors) when given a new AP trace of the validation population.

9.4.2 UKF Method

After formulating the estimation problem as a non-linear state-space representation where a noisy voltage trace is considered as the observed variable and SDEs defining a ventricular human cell model are used to describe the process equations, the UKF method was applied for joint model parameter and state variables, providing not only mean estimates but also measurements of estimation uncertainty. The UKF algorithm presents better performance than other methods used for parameter and state estimation of non-linear state-space representations, such as the EKF algorithm, with the added advantage of not requiring calculation of Jacobians [JU04]. Also, as compared to other Monte-Carlo-based methods, such as Particle Filters [SCS12], the UKF algorithm is associated with notably lower computational costs.

When using the UKF, appropriate calibration of its hyperparameters σ_θ and σ_r , representing process and measurement variances, respectively, is a critical point to achieve high levels of accuracy. According to these previous results, an inadequate selection of these hyperparameters may lead to an increase in the estimation error above 50% of the value attained for optimally adjusted σ_θ and σ_r values. Based on a previous work [SPFBV+19], σ_r was set to $1mV$, equals to the variance of the measurement noise added to the clean synthetic AP signal. In the case of σ_θ , which is closely related to the convergence speed and potentially oscillatory behaviour of the estimates, its value was set to $\sigma_\theta = 10^{-8}$, as this value led to a minimum average mean absolute error in parameter estimation, as shown in Section 9.3.1. This value is in the range of feasible values shown in a previous work [SPFBV+19], with a slight difference in the optimal value justified by the fact that a higher number of model parameters were estimated in the present study as well as to the fact that a subunit-based formulation of SDEs for ionic gates, rather than the channel-based formulation used in [SPFBV+19], was here employed.

9.4.3 Combined DGDR-UKF Method by Initialisation and Updating

The use of DGDR estimates for both initialisation and updating of the UKF estimates has been demonstrated to play a very significant role in improving the estimation performance. On one hand, providing an initialisation for the UKF method in terms of its mean and covariance matrix based on DGDR estimates allowed reducing the mean estimation error and the uncertainty around it. Also, the convergence time was remarkably diminished, as described in Section 9.3.2. As a proof, the combined DGDR initialisation + UKF approach required approximately 35% of the number of beats than the individual UKF method to reach the same level of estimation uncertainty.

On the other hand, updating the UKF estimates at the end of each cardiac beat by using the DGDR estimates allowed the solution of the combined method to remain within a relatively narrow range around the actual parameter values and avoided the estimation to fall into local minima. In addition, it contributed to accelerating estimation convergence, reducing by more than 95% the number of beats required by the UKF method to reach the same level of uncertainty. It is interesting to highlight that this updating process improved the estimation of not only the mean and covariance of the model parameters, but also of all other model state variables.

The combined DGDR-UKF method provides relevant advantages as compared to other methods used in the literature for similar purposes. The DGDR-UKF method renders a one-to-one matching between input AP traces and the sets of estimated parameters, whereas other methods based on Genetic Algorithms, Moment-Matching or Gaussian Process Emulators provide only parameter estimates at a population level [LDC+18, TLRG17, CSC15]. In addition, when comparing the DGDR-UKF method with other methods rendering individual parameter estimates, such as Markov Chain Monte Carlo (MCMC)-based methods [JCB+16, JBG+16], the DGDR-UKF method presents lower computational costs. Also, it is able to deal with beat-to-beat variability and to provide estimates of not only the model parameters but also of the hidden state variables, thus improving the global estimation accuracy.

The enhanced performance and reduced convergence time attained by the combined DGDR-UKF method are particularly relevant for subsequent studies aimed at investigating repolarisation variability from human ventricular experimental voltage traces, which are commonly of short duration.

9.4.4 Estimation of Ionic Current Conductances at Baseline

Eight ionic current conductances were estimated at baseline conditions, as variations in those conductances have been postulated to be major factors for spatial (cell-to-cell) AP variability [PCV+11, SKH+15, LFYC98, SBOW+14, LDC+18]. Other studies in the literature have addressed estimation of ionic current conductances, even if not in all cases for as many currents as in this work and not always considering temporal (beat-to-beat) AP changes but just focusing on a steady-state AP [TLRG17, JCB+16, SPFBV+19]. In the present study, stochastic ventricular human cell models accounting for temporal variability were developed to improve the estimation accuracy by considering dynamic

information additional to the static information commonly considered in the literature. The eight estimated model parameters were multiplying factors for the conductances of six major ionic currents (I_{Ks}, I_{Kr}, I_{to}, I_{CaL}, I_{K1}, I_{Na}) and the maximal values of I_{NaCa} and I_{NaK} with respect to their nominal values in the ORd model.

The least accurate ionic channel parameter estimations (irrespective of the tested methodologies) were obtained for θ_{Ks} in line with results reported in [SPFBV+19]. This can be due to the intrinsic characteristics of the ORd model, in which the I_{Ks} current has little influence on the AP, and consequently on AP-derived biomarkers, at baseline conditions. Other experimental and computational studies support this outcome regarding the limited influence of I_{Ks} on the AP shape and duration at baseline [XGP+13, JVB+05, OR12]. Since a wide range of θ_{Ks} values generate little differences in the corresponding AP traces, accurate identification becomes challenging. This issue is framed within the context of identifiability and observability and may be solved in future studies by complementing the estimation process with signals obtained while stimulating the cells at other pacing frequencies or under ionic current blocks. Similarly, the estimation errors associated with θ_{NaCa} and θ_{NaK} were among the highest for all tested methodologies, which can in this case be due to the longer time scale required for I_{NaCa} and I_{NaK} variations to impact the AP.

Of note, estimation of θ_{to} rendered much higher errors when the DGDR method was used as compared with any of the other methods involving UKF. This may be attributed to the fact that none of the defined AP-derived biomarkers may be closely related to the AP notch, which is the AP phase where this current has the largest influence. Similarly estimation errors of θ_{K1} were higher for DGDR than for any UKF-based method. In this case, despite considering biomarkers in the DGDR method like the resting membrane potential, which are expected to contribute to θ_{K1} identification, the UKF-based methods can deliver more accurate results because they use all samples of the AP trace, both during the AP as well as during the resting phase, and thus have a larger amount of information to adjust θ_{K1} estimation.

9.4.5 Estimation of Phosphorylation Levels of Cellular Substrates under $\beta-$AS Conditions

The phosphorylation levels corresponding to the three cellular substrates most significantly contributing to AP changes under $\beta-$AS were estimated using the proposed DGDR-UKF method and compared with other tested methods. To the best of our knowledge, this is the first study where the phosphorylation levels of a $\beta-$adrenergic signalling model have been estimated, together with other state variables, based on the static and dynamic AP changes induced by $\beta-$AS. The results obtained with the proposed combined method were generally better than those of individual DGDR and UKF methods. Nevertheless, it should be noted that the average mean absolute errors obtained for phosphorylation levels under $\beta-$AS were higher than those obtained for ionic conductances at baseline. This can be partly explained because the errors in the estimated baseline conductances were propagated to the estimation of the phosphorylation levels, as the latter were calculated based on the corresponding APs estimated at baseline.

While ionic conductances and phosphorylation levels under $\beta-$AS can be estimated simultaneously, it is degraded by the multiplicative relation of ionic conductances and phosphorylation levels in the coupled electrophysiological adrenergic signalling model. On the basis of such multiplicative relation, many combinations of conductance and phosphorylation level values could lead to the same estimation results even if the estimated parameter values were in fact far from their actual values.

9.4.6 Characterisation of Spatio-temporal AP Variability from Parameter Estimates

A main purpose of this study is to propose a method suitable for investigation of temporal and spatial variability in human ventricular repolarisation, with one-to-one identification of an underlying computational AP model for each experimentally available voltage trace. Provided data is available at baseline and under $\beta-$AS conditions, the proposed DGDR-UKF method can identify the specific electrophysiological and adrenergic signalling characteristics at those two conditions. This method was indeed able to precisely reproduce the AP shape, duration and variability of individual AP traces, rendering statistical distributions of the errors in the estimation of APD and STV remarkably more concentrated around 0 than those obtained with other tested methods, particularly when compared with the DGDR method.

On top of the DGDR-UKF method rendering better match between actual and estimated AP-derived biomarkers than other methods, it led to improved match between actual and estimated voltage traces, as illustrated in Section 9.3.6. This can be justified on the basis that this methodology provides estimates of not only the parameter values but of the complete vector of model state variables, which allows for more accurate AP reconstruction.

9.4.7 Limitations and Future Studies

In this work a total of 11 different human ventricular cell model parameters have been identified, corresponding to 8 ionic current conductances at baseline and 3 phosphorylation levels under $\beta-$AS. Future studies could include estimation of additional ionic current conductances (e.g. for I_{Cab}, I_{Nab}, I_{Kb} or I_{pCa}), phosphorylation levels (e.g. for ryanodine receptors, phospholamban or troponin I) or time constants of ionic gates (e.g. τ_{xrs}, τ_{xs1} or τ_{xk1}). Also, stochasticity could be added to other ionic currents like the late sodium current, which can have a relevant contribution to BVR.

To test the performance of the proposed methodology for estimation of model parameters and one-to-one replication of AP traces and AP-derived biomarkers, synthetic voltage traces were generated at $1Hz$ stimulation frequency. Future studies could test the extent to which the estimation performance is improved by applying the proposed DGDR-UKF method onto voltage traces obtained at different stimulation frequencies. In addition, voltage traces could be generated under different ionic blocks to offer additional information to be used for parameter identification, which could prove particularly useful

for identification of θ_{Ks}, θ_{NaCa}, θ_{NaK}, whose estimation was the most challenging in the present work.

A set of AP-derived biomarkers were used in the DGDR method and, consequently, in the DGDR-UKF method. Those biomarkers reflect AP characteristics related to its upstroke, repolarisation and resting potential as well as temporal APD variability. Novel AP-derived biomarkers reflecting additional information from the AP notch and plateau phases could help in the identification of model parameters, like θ_{to} and θ_{CaL} , thus globally improving the performance of the DGDR method and of the combined DGDR-UKF method.

This study has presented the combined DGDR-UKF method and has assessed its performance over a large set of synthetically generated AP traces. As a next step, the proposed method could be tested over experimental AP traces recorded from human ventricular cardiomyocytes or even extend the method to be applied onto voltage traces measured from ventricular human tissues. This would allow identification of underlying computational tissue models with representation of cell-to-cell electrical coupling.

9.5 Conclusion

A novel methodology based on the combined use of Double Greedy Dimension Reduction (DGDR), with Automatic Generation of Biomarkers, and the Unscented Kalman Filter (UKF) has been proposed to estimate parameters and state variables of an underlying human ventricular action potential (AP) model for any given input voltage trace. The proposed methodology is tested over synthetic voltage traces generated from an experimentally calibrated population of stochastic ventricular human cell models at baseline and under $\beta-$adrenergic stimulation. The combined methodology remarkably improves the estimation performance of individual DGDR and UKF methods while reducing the computational cost. The estimated ionic current conductances at baseline conditions and phosphorylation levels of cellular substrates under $\beta-$adrenergic stimulation allow for computational characterisation of spatio-temporal ventricular repolarisation, which can prove very useful to investigate variability changes induced by disease or drugs, uncover its underlying ionic mechanisms and establish a relationship with arrhythmia risk.

Chapter 10

Automated Patch-Clamp signal classification

Automated patch-clamp was developed to increase throughput and reduce the time consumption as compared to manual patch clamp, thereby making the technique feasible for large compound screens. Indeed, an automated patch-clamp device is able to patch hundreds of individual cells at the same time, allowing several compounds at different concentrations to be tested simultaneously (see Section 2.3.1.2). It follows that drug discovery might be highly accelerated. Several works on automated patch-clamp have already been done in safety pharmacology [BF21]. To pursue in this direction, it is then necessary to automise the analysis process, for instance by helping the experimenter to prune irrelevant compounds (within the meaning of a certain question described by the experimenter).

This work is devoted to the classification of compounds based on the Nav1.7 channel activity of human rhabdomyosarcoma muscle cells. This classification consists to detect whether a compound at a given concentration modulates the Nav1.7 channel. To investigate this, two classification strategies are used to answer the classification problem: a statistical evaluation directly based on biomarkers and the DGDR method based on the dictionary matrix extracted from the signals.

The following work was performed in collaboration with Kadla Røskva Rosholm, Lasse Homann and Anders Lindqvist, members of Sophion Bioscience[1].

[1]Ballerup, Denmark. sophion.com

Contents

10.1 Introduction

To preserve the benefit of automated patch-clamp, it is necessary to reduce the time of signal analysis. Indeed, the strategy is to construct a way to automatically detect relevant compounds (at a given concentration) with respect to a given question raised by the experimenter. Taking the benefit of it, the experimenter can go further in the investigation of retained samples.

In this context, we focused on one ion-channel the voltage-gated sodium channel: Nav1.7.

Nav1.7 channels are known to take part in the nociception. Its inhibitors are then studied to treat pain. However, several side effects may appear, showing the importance of a good dosage [KPCK20]. For these reasons, it is essential to study whether a compound at a given concentration blocks or modulates the Nav1.7 current.

This question falls into binary classification tasks with 'No Hit' (control-like) and 'Hit' (not control-like) as output labels. Some works have already been done on automated patch-clamp to study drug effects on sodium channels [CPZ+09, LBF11].

In this chapter we use the DGDR method on a dictionary entry constructed from automated patch-clamp signals in order to maximise the Hit/No Hit classification success rate on Nav1.7 ionic channels.

10.2 Material & Method

In this section, processes to implement the experiment, prepare data before the analyses and the post-processing are presented. The Experimental protocol section (see Section 10.2.1) describes the device, cells and voltage protocol to record patch-clamp signals. Then, in Section 10.2.2 the methodology to construct the dictionary entry is described. Finally, the Post-processing part (see Section 10.2.3) shows how samples are classified.

10.2.1 Experimental protocol

The experiments were performed using the automated patch clamp system Qube384. The measurement plate for this system is called the QChip and comprise 384 measurement wells each harbouring one cell for the analysis (see Figure 10.1[2]).

In each well, many cells are deposited, in such a way that one of them is placed on the electrode to perform the electrical activity recording.

Cells used for this study are human rhabdomyosarcoma muscle cells (RD)[3] provided by ATCC[4] and cultured in DMEM[5] supplemented 10% FBS penicillin/streptomycin

[2]Documentation and Figure are available on: https://sophion.com/products/qchips/.
[3]Cell line documentation available here.
[4]https://www.atcc.org/.
[5]* https://www.sigmaaldrich.com/.

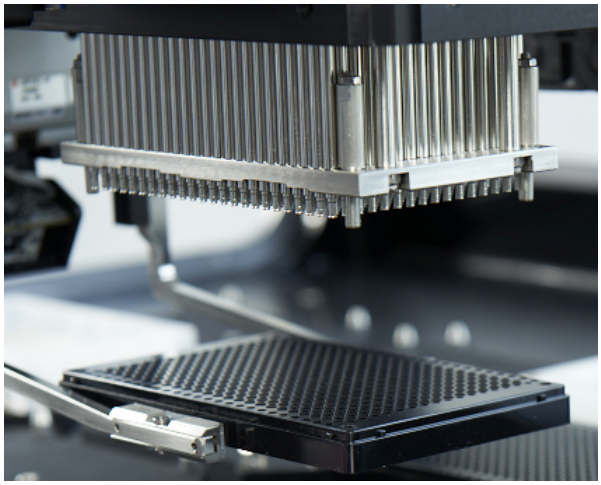

Figure 10.1: QChip 384 from Sophion Bioscience. The QChip corresponds to the black plate with its 384 wells (with electrodes to measure the electrical activity) in which each micropipette (top of the Figure) adds a compound at a given concentration. Used with permission from Sophion Bioscience.

$(100U/ml)$. Cells were harvested using Detachin[6] and stored in EX-CELL serum-free medium* supplemented with $25mM$ HEPES until experiment.

These cells have the main characteristic to endogenously expressing Nav1.7 ion channels. This expression allows to study its modulation by compounds. As this channel is highly expressed in nociceptors and being partly responsible for the pain, it has many interests in drug assessment to treat pain [CWA+16, KPCK20].

As non-pacemaker cells, a voltage protocol is applied at control case (without compound) and after compound addition. It consists in imposing various voltages at different times and recording the current response from the cell (see Figure 10.2 for the voltage protocol).

Each step to $-10mV$ from a holding potential of $-100mV$ induces sodium channel opening leading to a sodium spike current.

A sweep corresponds to the recorded current in response to one application of the full voltage protocol, which is then repeated 5 times at each condition (control case and compound case). Experiment being performed on 10 QChip, it follows that a total of $n_s = 19200$ samples are available for the study.

10.2.1.1 Compounds

Among the $n_s = 19200$ samples, a total of 7 compounds at 4 different concentrations were considered in this study. Additionally, Tetrodotoxin at $1\mu M$ was used as a positive control (state-dependent sodium blocker). Cisapride is the only tested compound without

[6] http://www.genlantis.com/detachin.html

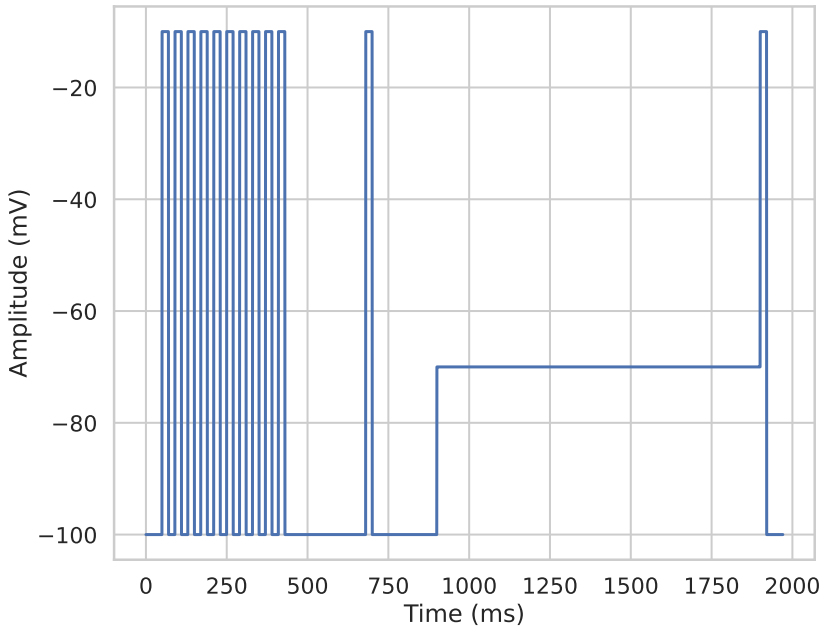

Figure 10.2: Voltage protocol.

blockade action on the sodium channel (non-active). All other blockers have a state-dependent mechanism except Anemone toxin II which has a delay inactivation effect. A summary is given in Table 10.1.

Compound	Abbreviation	C1	C2	C3	C4	Mechanism
Amitriptyline	ATT	10	1	0.1	0.01	State-dependent
Anemone toxin II	ATXII	0.01	10^{-3}	10^{-4}	10^{-5}	Delays inactivation
Bupivacaine	BPV	30	3	0.3	0.03	State-dependent
Cisapride	CSP	5	0.5	0.05	0.005	Non-active
Flecainide	FCN	30	3	0.3	0.03	State-dependent
Mexiletine	MXT	100	10	1	0.1	State-dependent
Tetracaine	TRC	30	3	0.3	0.03	State-dependent
Tetrodotoxin	TTX	1	-	-	-	State-dependent

Table 10.1: Compounds and concentrations used for the study. Concentrations are in μM. Mechanism of action on the Nav1.7 channel.

A part of the experimental dataset was performed with 0.3% DMSO as negative controls which are either tagged 'Blank' or 'Neg'.

10.2.1.2 Signal traces

Some examples are shown in Figure 10.3. In the case of 'Hit' and 'No Hit' binary classification, the goal is to detect whether a compound at a given concentration modulates the signal or not, based on traces recorded such as in Figure 10.3.

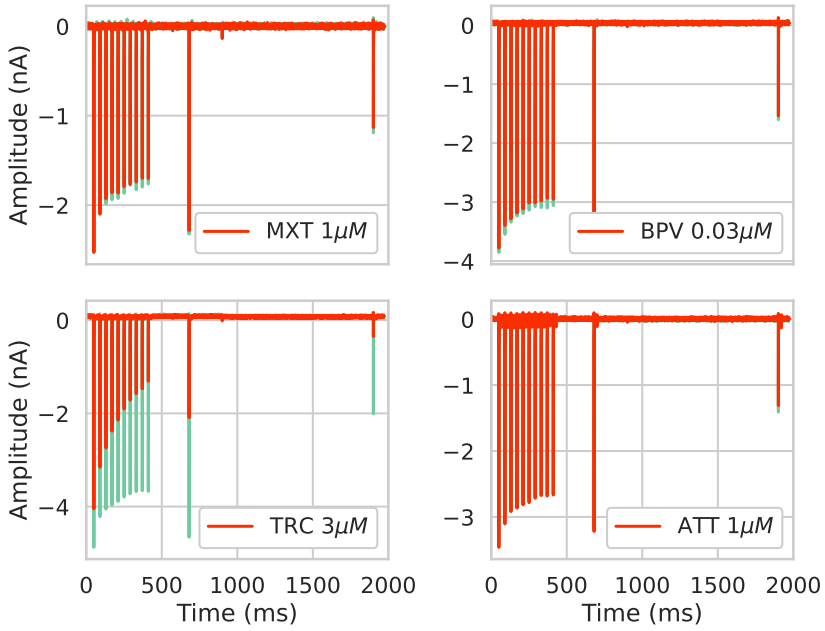

Figure 10.3: Examples of sweeps before (control: blue traces) and after (drug: orange traces) compound addition.

To construct the classifier based on the DGDR method, we will first need to construct a dictionary from these signals.

10.2.2 Pre-processing

In order to apply the DGDR method described at the beginning of the manuscript, we need to extract several quantities to construct the dictionary entry set. The extraction is performed for each sample, at control case and under compound addition. Once the dictionary set is constructed, it has to be split into three sets: Training, Validation and Test set. Finally, a rescaling is performed to improve KNN performances.

10.2.2.1 Dictionary entry computations

First, each sweep is split into 12 traces (one for each spike). Then, for each trace, 6 quantities are computed: amplitude, spike area, electric charge (see original definition in

Definition 14) at 25%, 50%, 75% and 100% of the trace period (see Figure 10.4 for the amplitude and electrical charge at 50% of the trace period).

Definition 14
Let $I(t)$ be the current at time t and t_1 and t_2 two times such that $t_2 > t_1$. The electrical charge Q between t_1 and t_2 is given by:

$$Q = \int_{t_1}^{t_2} I(t)dt.$$

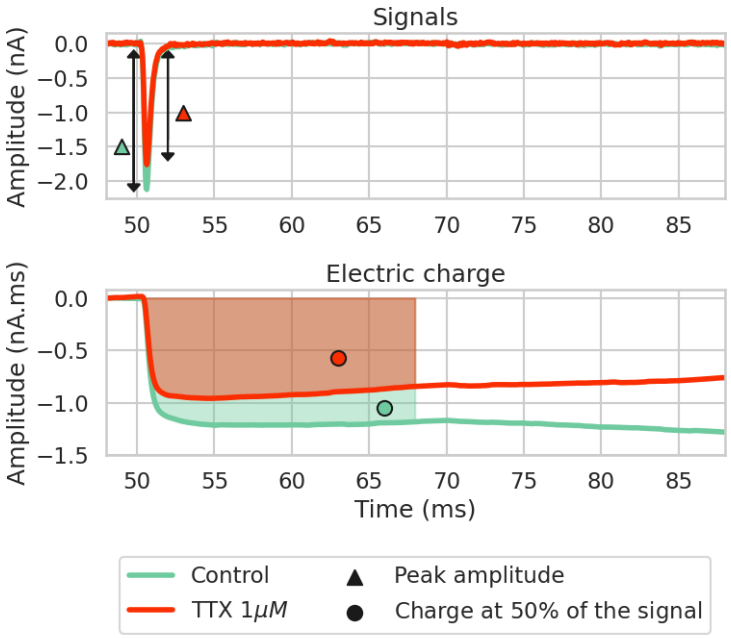

Figure 10.4: Example of amplitude and electrical charge at 50% of the spike trace at baseline case and under addition of $1\mu M$ of TTX (control positive).

In complement, two by two cross products are considered to introduce non-linearities. It results in 27 quantities extracted per beat. These quantities are computed at control case and under compound addition. We then consider the ratio between compound entries and corresponding entries for the control case (in the same well). There are two main reasons to justify this choice. On one hand, it exists a cell variability between wells. Then, signals may be different even if the compound and its concentration are the same in two different wells. However behaviours with respect to its control cases should be closer. On the other hand, we are considering a Hit/No Hit classification. Using relative values between control and control (e.g. before addition and without or after negative control addition) allows us to consider the 'No Hit' variability.

As a preliminary study, for the first study case (see its definition in Section 10.2.2.2 and results in Section 10.3.1) wavelet coefficients were added to enrich the dictionary. These coefficients are computed on the absolute difference of each beat and on the whole sweep (absolute difference between the under compound addition case with its corresponding control case). The strategy used was the same a the one described in Section 13.2.3.2. It led to a dictionary entry matrix of size $n_g = 3782$.

For other studies (see its definitions in Section 10.2.2.2 and results in Section 10.3.2), the same quantity of each beat is averaged (over the 12 beats) and considered into the dictionary. It follows that, for these studies, the dictionary entry size is $n_g = 27$.

10.2.2.2 Sets generation

Once data are collected, a random splitting is performed into three sets in such a way that each sample belongs to one and only one set. These three sets are the Training set Tr, the Validation set Va and the Test set Te. The above constraint on each sample allows us to use all of them and mostly to avoid having a same sample in the different sets (which can be seen as an inverse crime). In this chapter, 4 studies were performed following the protocol below:

1. The first test can be seen as a control test and refers to Section 10.3.1. It consists in the 'Hit' impact detection of each compound at each concentration.
    - Number of repetitions: 10.
    - Dictionary entry size: $n_g = 3782$.
    - Training set of size 40 (20 samples labelled 'No Hit' and 20 samples labelled 'Hit').
    - Validation set of size 40 (20 samples labelled 'No Hit' and 20 samples labelled 'Hit').
    - Test set: Unused elements for the 'Hit' part and same number for the 'No Hit' part. Despite each protocol is repeated 5 times, the number of experiments for each compound at a given concentration is not the same. For this reason the test set size is not always the same. For this study the average test size is about 120. The lowest test size is 70 whereas the highest is 170.
2. The second test (see Section 10.6.0.1), referred to as (800) in what follows, consists in dividing the whole set of experiments into three parts:
    - Number of repetitions: 100.
    - Dictionary entry size: $n_g = 27$.
    - Training set of size 800 (400 samples labelled 'No Hit' and 400 samples labelled 'Hit').
    - Validation set of size 800 (400 samples labelled 'No Hit' and 400 samples labelled 'Hit').

- Test set of size 17600 (15155 samples labelled 'No Hit' and 2445 samples labelled 'Hit').

In particular, for training and validation sets, we considered as 'No Hit': CSP, Blank and Neg experimental samples and as 'Hit': ATT, ATXII, BPV, FCN, MXT, TRC and TTX for all concentrations.

3. The third test (see Section 10.6.0.2), referred to as (C3C4-800) consists in dividing the whole dataset as follows:

   - Number of repetitions: 100.
   - Dictionary entry size: $n_g = 27$.
   - Training set of size 800 (400 samples labelled 'No Hit' and 400 samples labelled 'Hit').
   - Validation set of size 800 (400 samples labelled 'No Hit' and 400 samples labelled 'Hit').
   - Test set of size 17600 (16450 samples labelled 'No Hit' and 1150 samples labelled 'Hit').

   In particular, for training and validation sets, we considered as 'No Hit': CSP, Blank, Neg, all the molecules at concentration C4, all the molecules at concentration C3 except TRC, and ATT at concentration C2. We considered as 'Hit' all the rest. A summary is given in Table 10.2.

4. The fourth test (see Section 10.3.2.3), referred to as (1600) consists in dividing the whole dataset as follows:

   - Number of repetitions: 100.
   - Dictionary entry size: $n_g = 27$.
   - Training set of size 1600 (800 samples labelled 'No Hit' and 800 samples labelled 'Hit').
   - Validation set of size 1600 (800 samples labelled 'No Hit' and 800 samples labelled 'Hit').
   - Test set of size 16000 (15650 samples labelled 'No Hit' and 350 samples labelled 'Hit').

   In particular, for training and validation sets, we considered as 'No Hit': CSP, Blank, Neg, all the molecules at concentration C4, all the molecules at concentration C3 except TRC, and ATT at concentration C2. We considered as 'Hit' all the rest. A summary is given in Table 10.2.

Remark 15
For each study cases above, a sample belongs to one and only one set (i.e. Training, Validation or Test set).

Compound	C1	C2	C3	C4
ATT	Hit	No Hit	No Hit	No Hit
ATXII	Hit	Hit	No Hit	No Hit
BPV	Hit	Hit	No Hit	No Hit
CSP	No Hit	No Hit	No Hit	No Hit
FCN	Hit	Hit	No Hit	No Hit
MXT	Hit	Hit	No Hit	No Hit
TRC	Hit	Hit	Hit	No Hit

Table 10.2: Label given for each compound at each concentration for study cases 3 ad 4.

Remark 16
*The Training set size being much lower than the Test set size for study cases two to four, a higher number of repetitions (*100*) was performed for the majority vote strategy. To hilight the convergence, for each repetition n, we compute the normalised Hamming distance between the classification output obtained through the majority vote up to $n-1$ repetitions and up to n repetitions. In particular, for the study number four, it led to a normalised Hamming distance lower than 7.10^{-4} for the last repetition (i.e normalised Hamming distance between the classification output over the first* 99 *repetitions and the classification output over the* 100 *repetitions). It means that, over the* 19600 *samples, around* 14 *do not have the same output label between two consecutive repetitions. The convergence for the study case number four is shown in Figure 10.12 in the Appendix.*

10.2.2.3 Data Rescaling

Once entries are computed they may fall in a high value range. This might be binding, particularly when using a KNN methods. To avoid this, we suggest rescaling each entry into a unit hypercube. In particular, it would lead to an easier comprehension of the selected entries and their weights. Usually, we compute the minimal/maximal value of each entry of the Training set Tr and proceed as follows for the rescaling:

$$\begin{cases} M_i = \max(Tr_i), m_i = \min(Tr_i) \\ \begin{cases} Tr_i = \frac{Tr_i - m_i}{M_i - m_i} \\ Va_i = \frac{Va_i - m_i}{M_i - m_i} \\ Te_i = \frac{Te_i - m_i}{M_i - m_i} \end{cases} \end{cases},$$

where i denotes the i^{th} dictionary entry, Tr is the Training set, Va the Validation set and Te the Test set. However, it may appear that in some cases, some outliers are inside the Training set. It follows that the Training set is in the unit hypercube, but for some directions (i.e. dictionary entries), data are condensed. This would affect the process of entry selection (based on a KNN approach) and afterwards the classification. To overcome this, we suggest the following approach:

$$\begin{cases} \mu_i^{(0)} = \mathbb{E}[Tr_i^{(0)}], \mu_i^{(1)} = \mathbb{E}[Tr_i^{(1)}] \\ \sigma_i^{(0)} = \sigma(Tr_i^{(0)}), \sigma_i^{(1)} = \sigma(Tr_i^{(1)}) \\ M_i = \max(\mu_i^{(0)} + 2\sigma_i^{(0)}, \mu_i^{(1)} + 2\sigma_i^{(1)}), m_i = \min(\mu_i^{(0)} - 2\sigma_i^{(0)}, \mu_i^{(1)} - 2\sigma_i^{(1)}) \\ \begin{cases} Tr_i = \frac{Tr_i - m_i}{M_i - m_i} \\ Va_i = \frac{Va_i - m_i}{M_i - m_i} \\ Te_i = \frac{Te_i - m_i}{M_i - m_i} \end{cases} \end{cases},$$

where $S_i^{(j)}$ stands for the i^{th} dictionary of set $S \in \{Tr, Va, Te\}$ restricted to class $j \in \{0,1\}$, $\mathbb{E}[.]$ and $\sigma(.)$ denote the empirical mean and standard deviation of the considered set respectively. This approach leads to a more robust rescaling in the face of the outliers.

Remark 17
The above approach is described for a binary classification problem, but can trivially be extended to multiclass problems.

10.2.3 Post-processing

As the classification is repeated 100 times, the given label for each sample in the Test set is the majority voting strategy [Nar05] described in Algorithm 6.

Algorithm 6 Majority voting.

Require: M {Output label for each sample, each time selected to be into the Test set.}
  $y \leftarrow 0$ {Initialise output label vector $y \in \mathbb{R}^{n_s}$ to 0.}
  for M_i, y_i **do**
    $c_i \leftarrow counter(M_i)$ {Count the number of times sample i is classified in each class.}
    $y_i \leftarrow argmax(c_i)$ {Final label using majority voting on sample i.}
  end for
  return y

10.3 Results

This section is divided into two parts. The first part corresponds to the first study case described in Section 10.2.2.2. The second part is devoted to the three other study cases. All the following studies were performed with 5 nearest neighbours and *a priori* $\pi_0 = \pi_1 = \frac{1}{2}$.

10.3.1 First study case: Validation of the method

Results are summarised in Figure 10.5.

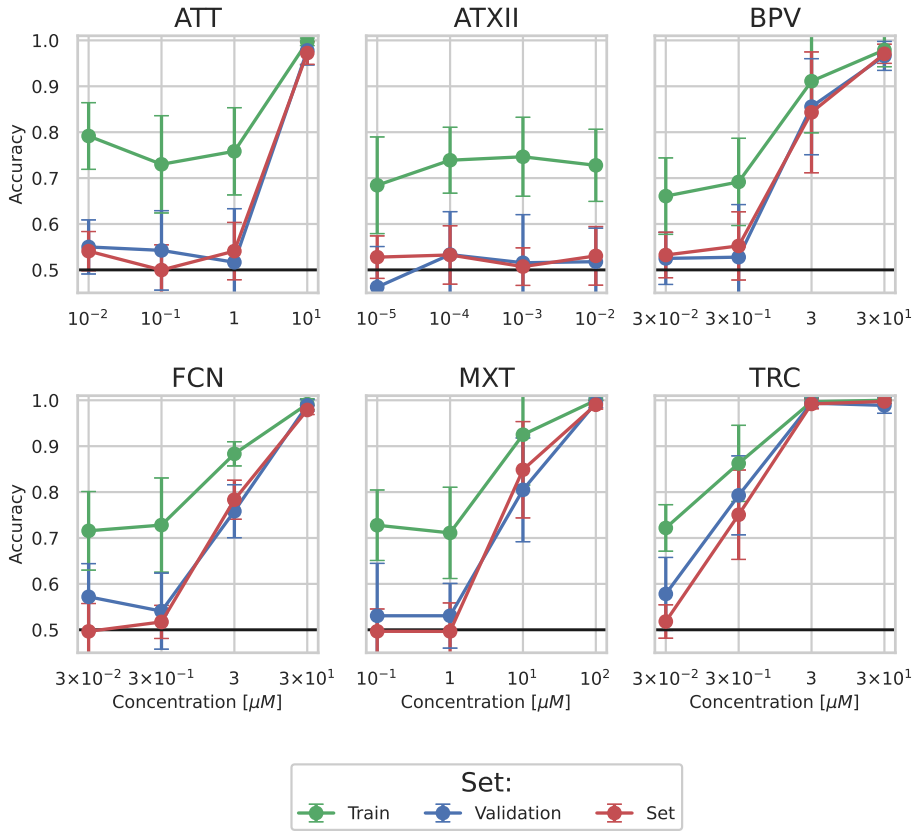

Figure 10.5: First study case: Accuracies obtained for each compound at a given concentration.

The Training set always shows higher accuracy as the DGDR method was performed on the Training set (inducing a slight overfitting). The Validation set was used to set a stopping criterion (early stopping). The Test set was not used for the learning process. The fact that Validation and Test sets show similar results is then expected (this is to show that the random construction of the sets works).

10.3.1.1 Detailed results:

- State-dependent blockers: For this mechanism, we clearly see that the classification success rate (accuracy) increases with the concentration. In particular, at low concentration, the discrimination between 'No Hit' and 'Hit' cannot be well performed. In particular, MXT at $1\mu M$ and BPV at $0.03\mu M$ shown in Figure 10.3 fall into the control case variability. Conversely, at higher concentrations, we are almost sure to always detect a hit. We can also see that the overfitting on the training set tends to decrease as the concentration increases.

- Delays inactivation blocker (ATXII): Accuracies for validation and test sets are

closed to 0.5 regardless of the concentration used. It means that we cannot make the discrimination between the negative control sample and ATXII. However, ATXII being deactivated during the course of the experiments, it is then expected to not detect an effect on these recorded signals.

- Assumptions on the $IC50$: Tetracaine (TRC) shows good results, even at low concentration ($0.3\mu M$) which is not the case for the other compounds (and particularly for those who have the same concentrations: BPV and FCN). The same scenario appears for BPV with respect to FCN (higher accuracy for the same concentration). Denoting by $IC50(X)$ the $IC50$ of the compound X to the Nav1.7 channel, we can make the following assumption: $IC50(TRC) < IC50(BPV) < IC50(FCN)$.

The case where the Hit/No Hit classification does not lead to a good classification score is induced by the fact that there are no relevant entries to solve the classification task. In particular, from the 10 runs, a higher variability on the selected entries appears. This scenario is observed for ATXII and compounds at low concentration. Most selected entries (for the two highest concentrations) are given in Table 10.3.

Name	Frequency (%)
$Area_D * Ch_{50}$ of beat 3	23
Amp of beat 0	19
$Area_D$ of beat 7	13

Table 10.3: First study case: Frequencies ($\geq 10\%$) of selected entries over the 10 runs. Ch_X denotes the electric charge at $X\%$ of the beat period, Amp the maximal amplitude of the beat and $Area_D$ the surface of the cell depolarisation.

10.3.2 General Application

In the previous section, results show that the Hit/No Hit classification for a given compound and concentration works well, particularly for higher concentrations. This preliminary study is crucial for the Hit/No Hit classification. Indeed, even if for the highest concentrations we had bad results, we would have no hope to get good results in a more general case (mixing compounds and concentrations in the Training, Validation and Test sets). These encouraging results, obtained in the First study case, allow us to extend Hit/No Hit classification studies into more general cases (i.e. studies two to four).

10.3.2.1 Computational Time

For the study case that took more time, around 8 minutes are needed for the training phase, which has been run in parallel on 27 physical CPUs. The testing phase took more or less a couple of seconds. Repeating this process 100 times, in order to have some statistics, took 13 hours for the training phase and few minutes for the tests. This is typical of this kind of strategy, in which we decompose the problem solving into two

parts: an offline phase, which is here the training phase, performed once and for all, in which we concentrate all the computational efforts; an online phase, which is here the test phase, which benefits from the learning phase and that can be performed in few seconds. Remark that, given new data in the same experimental setting, there is no need to re-run the learning phase.

10.3.2.2 Second and Third study cases

For sake of clarity detailed results of these two study cases are described in the Appendix (see Sections 10.6.0.1 and 10.6.0.2 for study cases two and three respectively). As defined in Section 10.2.2.2, the second study does not take into account results obtained in the first study, meaning that for each compound, the same label is given regardless of the considered concentration. Despite this choice, the classification success is quite good, being close to 0.86. To consider the fact that we are not able to distinguish well 'Hit' and 'No Hit' samples for the lowest concentrations, the third study was established (see Section 10.2.2.2 for details on the set generation). For this scenario, the accuracy reached to 0.95. The fourth study was established to consider a higher Training and Validation sample size and a more robust classifier. It led to similar results as shown in the following section.

10.3.2.3 Fourth study case

The confusion matrix for this test case is shown in Figure 10.6.

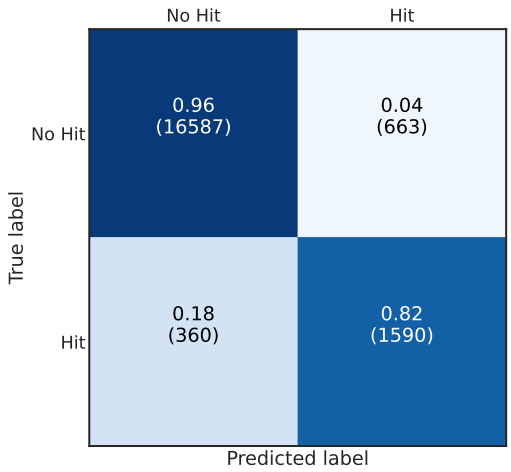

Figure 10.6: Fourth study case: confusion matrix obtained for Hit/No Hit classification.

We see that the accuracy rate improves and it reaches 0.95. We improve as well false positive and negative. Remark that these numbers refer to the case in which we make the

hypothesis that at the lowest concentration the molecules behave as 'No Hit'. Cohen's kappa is around 0.73.

Detailed results: Here are presented classification results for positive Figure 10.7 and negative Figure 10.8 compounds at each concentration. All compounds are well classified for its higher concentration (except the ATXII as a deactivated compound).

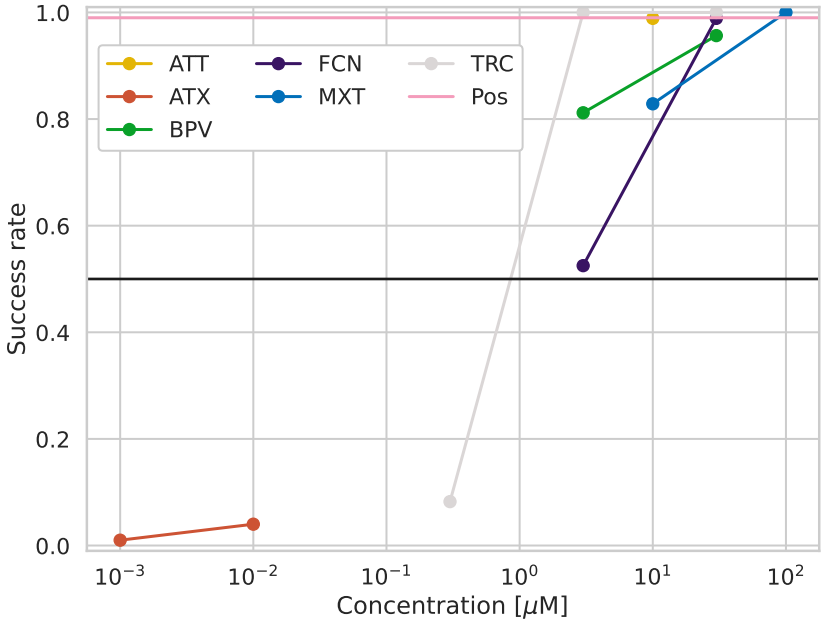

Figure 10.7: Fourth study case: Classification details for 'Hit' compounds. 'Pos' stands for the positive control: Tetrodotoxin at $1\mu M$.

TRC is still not that well classified at C3. We see that in this configuration we improve the confidence of the classification (meaning that we classify as 'No Hit' the lowest concentrations and as 'Hit' the largest, with more certainty). Remark that the Figures from test cases 3 and 4 are quite similar. This confirms that the approach is quite robust. Most selected entries are given in Table 10.4.

Remark 18
The most selected entries are essentially the same as for the Third study case (see Table 10.8), which highlight the robustness of the classifier.

10.3.2.4 Comparisons

In this section, we compare results obtained using the DGDR method and the method currently used at Sophion (S.E for statistical evaluation).

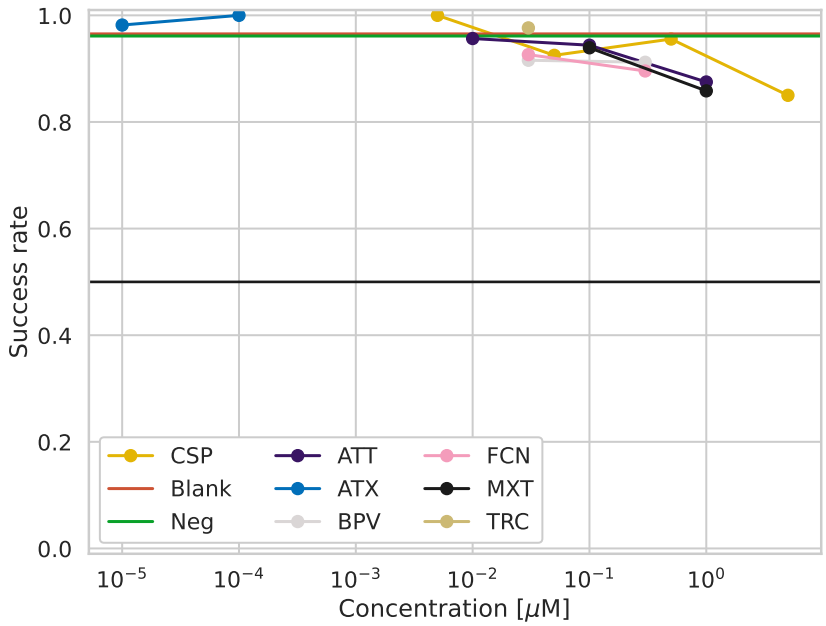

Figure 10.8: Fourth study case: Classification details for 'No Hit' compounds.

Name	Frequency (%)
Amp	26
Ch_{50}	16.3
$Ch_{25} * Ch_{50}$	15.7
$Amp * Ch_{50}$	11.1

Table 10.4: Fourth study case: Frequencies ($\geq 10\%$) of selected entries over the 100 runs. Ch_X denotes the average electric charge at $X\%$ of the beat period and Amp the average maximal amplitude of the sweep.

Remark 19

For the DGDR method, we considered classifiers obtained with the fourth study. The label given to a sample is determined by the use of the majority vote strategy described in Section 10.2.3. However, to make the comparison with the S.E method possible, the same True label of a tested sample is considered for the two methods. This means that we do not consider the assumptions made in the fourth study for the tested samples. This, explains why the accuracy obtained with the DGDR method is not the same in the following results.

The comparison of the two classifiers is given in Table 10.5.

Compounds are more classified as 'Hit' with the DGDR method than the S.E method.

		DGDR	
		No Hit	Hit
S.E	No Hit	2865	94
	Hit	56	126

Table 10.5: Comparison between the 'statistical evaluation' (S.E) proposed by Sophion and the DGDR method.

The two strategies are in agreement for approximately 95% of the elements of the whole sample. Details for each compound are shown in Figure 10.9.

The success rate obtained with the S.E method is around 0.893 whereas it is around 0.894 for the DGDR method. Despite this gain seems very low, data classes are highly unbalanced (17% of the data have a true label 'Hit'). The DGDR tends to more easily classify a compound has 'Hit', resulting in a lower specificity (0.97 versus 0.98 for the S.E method). However, the gain is higher on the sensitivity (0.52 versus 0.46 for the S.E method). These results are summarised in Table 10.6.

Quantity	DGDR	S.E
Accuracy	0.894	0.893
Specificity	0.970	0.980
Sensitivity	0.521	0.465
Precision	0.779	0.825
F1-score	0.624	0.594
Cohen's kappa	0.565	0.538

Table 10.6: Confusion matrices quantities.

Among these quantities only the specificity and precision are lower with the DGDR method. The gain on the sensitivity implies that we better classify 'Hit' compounds with the DGDR than with the S.E method. A summary for each 'Hit' compounds at a given concentration is summarised in Figure 10.9.

The DGDR globally leads to a higher accuracy on 'Hit' compounds than S.E method (even for the ATXII which should have not be detected as a deactivated compound). Moreover, the accuracy increases with the concentration, which is not the case for the S.E method (ATT, FCN and BPV at lower concentrations). The gain is particularly significant for the second-highest concentrations (BPV, FCN and MXT).

For the Cisapride case (CSP, which is a negative control), results are given in Figure 10.10.

As a negative control, we obtain a less good accuracy using the DGDR method than the S.E method. Several scenarios can explain these results. The first is the unbalanced dataset. The proportion of 'Hit' data is much lower than the 'No Hit' data. The *a priori* being set to 0.5 each, it explains why we tend to more classify compounds as 'Hit'. The second reason is the data preprocessing. Contrary to the S.E method, all the data were considered to construct the dictionary, meaning that some of them may pollute the

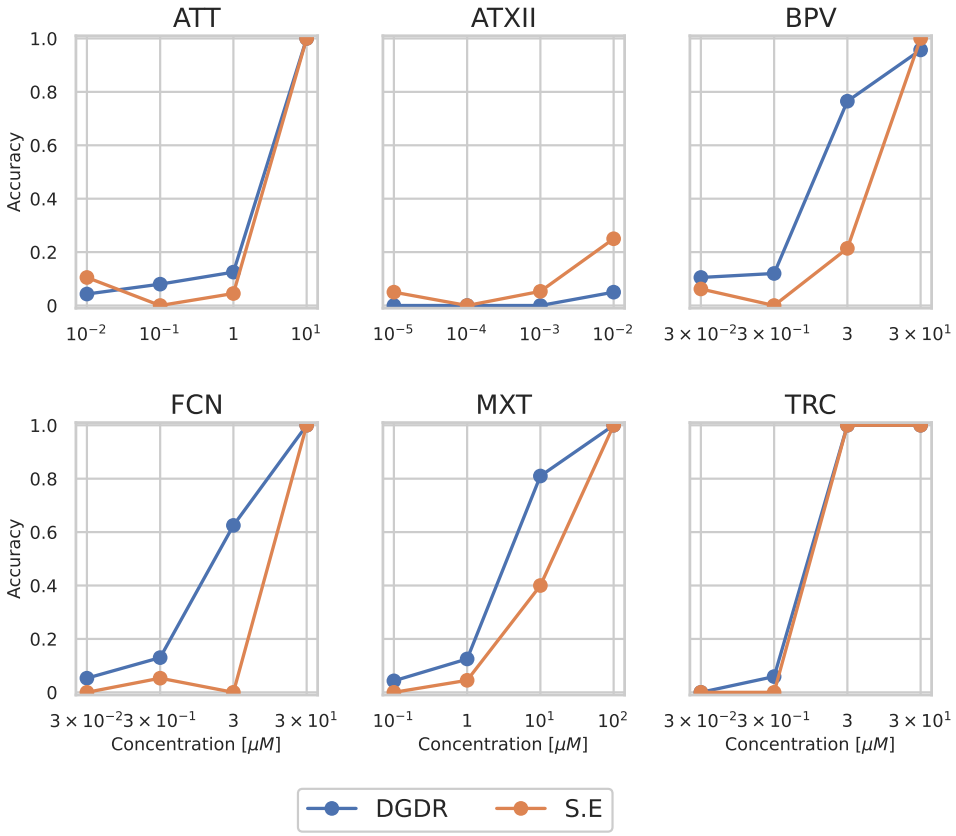

Figure 10.9: Classification success rate comparisons between DGDR and S.E methods for each positive compounds at a given concentration.

DGDR process. Further investigations could be made by combining the DGDR method with the ASE-HD method to construct an augmented training set. Another explanation concerning the Cisapride, could be that even if it is a negative control, it may affect the cell (by other mechanisms). The DGDR method may have selected entries for which this mechanism can be detected. This could explain why the accuracy tends to decrease as the concentration increases. A preliminary test would be to classify Cisapride as 'Hit' against control (no compound addition) as 'No hit'.

For each concentration (from C4 to C1, C4 being the lowest and C1 the highest), 6 quantities were computed from the confusion matrices obtained for the two methods: accuracy, specificity, sensitivity, precision, F1-score and Cohen's. The comparison of the two methods are shown in Figure 10.11.

As a deactivated compound, ATXII was tagged 'No Hit' to compute the confusion matrices and the resulting extracted quantities. Despite the two classification strategies show similar quantities at lower and higher concentrations, this is clearly not the case for in between concentrations. In particular, all the extracted indicators are higher for the

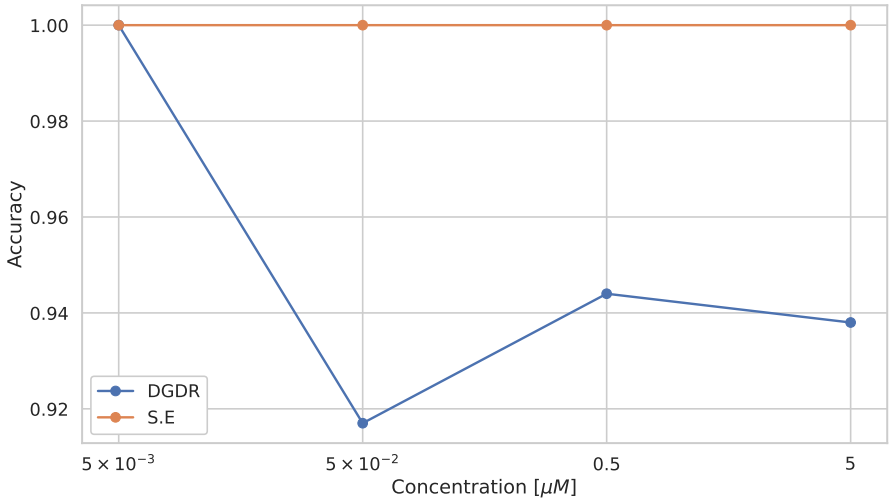

Figure 10.10: Classification success rate comparisons between DGDR and S.E methods for Cisapride at a given concentration.

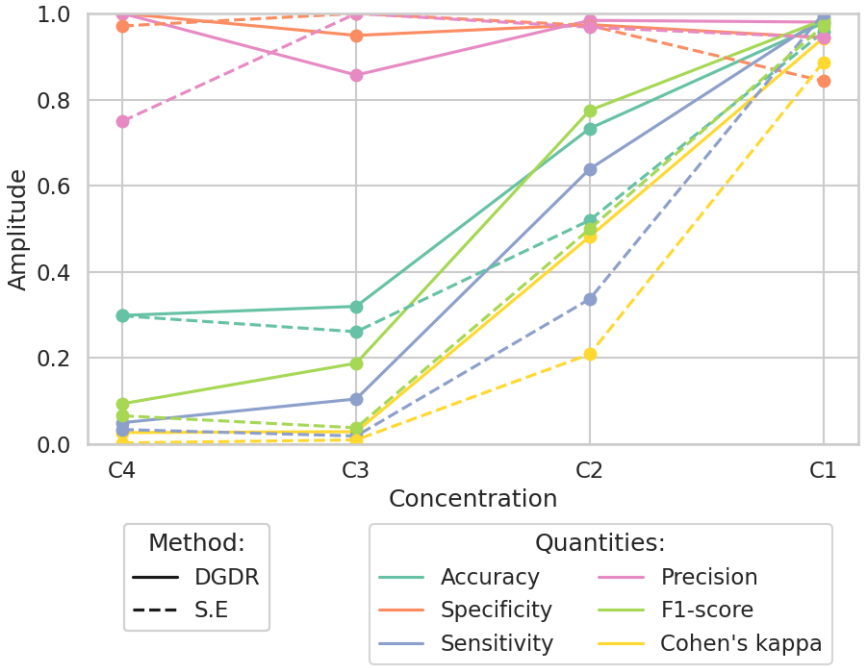

Figure 10.11: Comparison of the DGDR and S.E method using classical indicators. For this comparison, ATXII was considered as a 'No Hit' compound.

DGDR method at concentration C2 and concentration C3 (except for the specificity and precision).

10.4 Discussion

Studies two to four were performed on a restricted dictionary entry size (27 instead of 3782 for the first study case). Selected entries in study case one were mainly limited to agnostic entries (i.e. amplitude and electrical charge). Due to computational limits, only 27 entries (corresponding to average agnostic quantities over a sweep) were kept. Indeed, the process was repeated 100 before performing the majority vote. However, results are quantitatively good. And by construction of the DGDR method, results obtained using the largest dictionary entry size would be at least as good as results obtained in studies two to four.

10.5 Conclusion

The DGDR method applied to automated patch-clamp for 'Hit/No Hit' classification of sodium blockers seems to be quite efficient. In particular, as expected, the classification success rate increases with the concentration. Regarding to study one, the behaviour of each compound with respect to the concentration seems to allow us to order them in terms of $IC50$. Despite ATXII is badly classified, the obtained classification results are expected by the experimental setup (deactivated compound). All other compounds are almost perfectly classifiable for its highest concentrations. Moreover, as the overfitting decreases as the concentration increases, it means that selected entries are more and more relevant and robust to the classification task. These selected entries are mostly agnostic entries, such as amplitude or electric charge.

Studies two to four reach to a success rate close to 0.95 according to the label hypothesis made. These results are in a good agreement with those obtained in study case one. Despite the gain on the accuracy is weak in comparison with the statistical evaluation strategy considered at Sophion, the sensitivity increases from 0.465 with the statistical evaluation to 0.521 with the DGDR method. The quantities extracted from confusion matrices highlight that the DGDR method is more suitable than the statistical evaluation, particularly at intermediate concentrations.

Contrary to the statistical evaluation method, the DGDR method considered all the data without any pruning for preprocessing. An application of the ASE-HD method to construct the augmented training set could improve the success rate with the DGDR method.

The selected entries by the DGDR method allows a classification in agreement with the concentration (i.e the classification success rate for 'Hit' compounds increases as the concentration increases). This phenomenon is less clear (at lower concentrations) for the statistical evaluation strategy.

Further investigations could be done. In particular, the ASE-HD method could be used to construct an augmented training set and remove irrelevant data. Fewer compounds/concentrations could therefore be used to design experiments.

A 'Hit/No Hit' classification on Cisapride could be done to ensure that recorded signals (under Cisapride addition) does not carry this information as a non-active compound.

10.6 Appendix

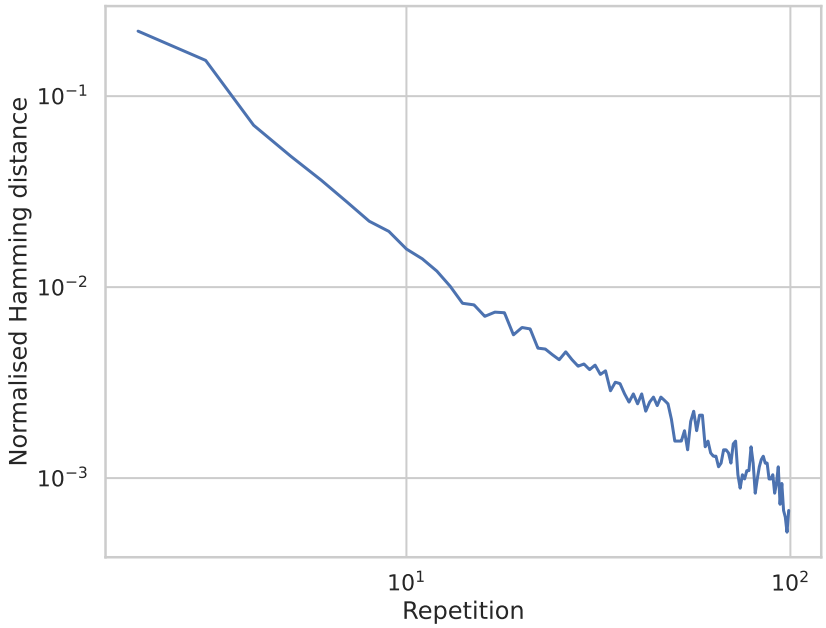

Figure 10.12: Study case four: convergence of the normalized Hamming distance between two consecutive output classification.

10.6.0.1 Second study case

Obtained confusion matrix is shown in Figure 10.13.

The accuracy is around 0.86 and Cohen's kappa is around 0.51. Note that the fact that the two classes are highly unbalanced may affect Cohen's kappa and then its interpretation. Here 0.86 is computed by considering the true labels as the one used to construct training and validation. To better understand false positives and false negatives, more details are given in the following section.

Detailed results: Here are presented classification results for positive (see Figure 10.14) and negative (see Figure 10.15) compounds at each concentration.

Except for ATXII, all compounds are well classified for its higher concentration. In particular, the classification success rate for TRC is higher than 0.5 as of the second concentration. All negative compounds have a good classification success rate, except for the highest concentration of Cisapride which is more or less equivalent to flip coin. Most selected entries are given in Table 10.7.

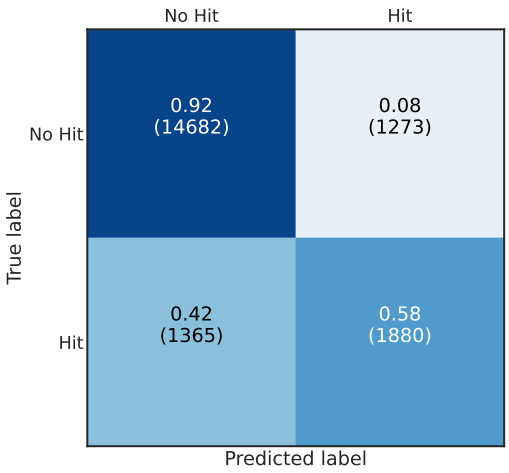

Figure 10.13: Second study case: confusion matrix obtained for Hit/No Hit classification.

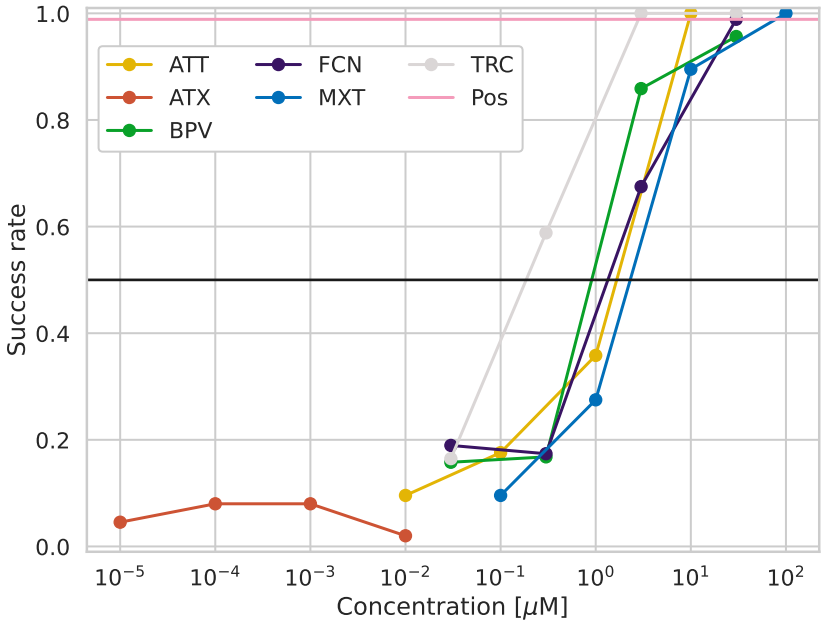

Figure 10.14: Second study case: Classification details for 'Hit' compounds.

10.6.0.2 Third study case

The confusion matrix for this test case is shown in Figure 10.16.
We see that the accuracy rate improves and it reaches 0.95. We improve as well false

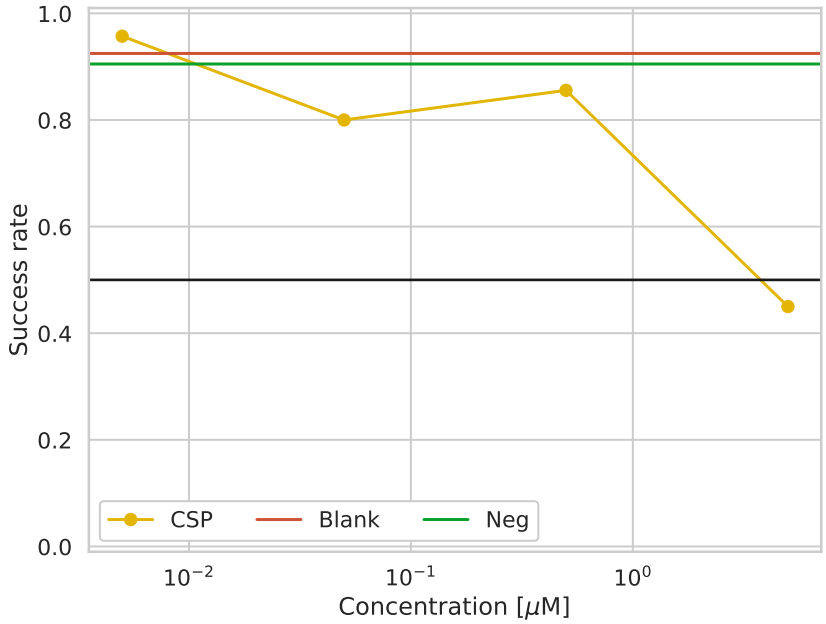

Figure 10.15: Second study case: Classification details for 'No Hit' compounds.

Name	Frequency (%)
$Ch_{25} * Ch_{50}$	18
Amp	11
Ch_{50}	11

Table 10.7: Second study case: Frequencies ($\geq 10\%$) of selected entries over the 100 runs. Ch_X denotes the average electric charge at $X\%$ of the beat period and Amp the average maximal amplitude of the sweep.

positive and negative. Remark that these numbers refer to the case in which we make the hypothesis that at the lowest concentration the molecules behave as 'No Hit'. Cohen's kappa is around 0.73.

Detailed results: Here are presented classification results for positive Figure 10.17 and negative Figure 10.18 compounds at each concentration.

All compounds are well classified for its higher concentration (execpt for ATXII as a deactivated compound). TRC is not that well classified at C3. We see that in this configuration we improve the confidence of the classification (meaning that we classify as 'No Hit' the lowest concentrations and as 'Hit' the largest, with more certainty). Most selected entries are given in Table 10.8.

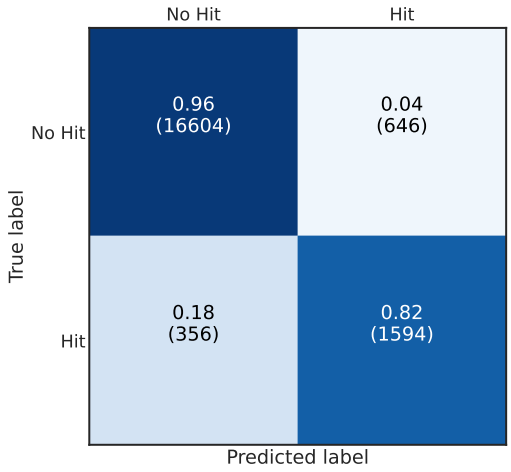

Figure 10.16: Third study case: confusion matrix obtained for Hit/No Hit classification.

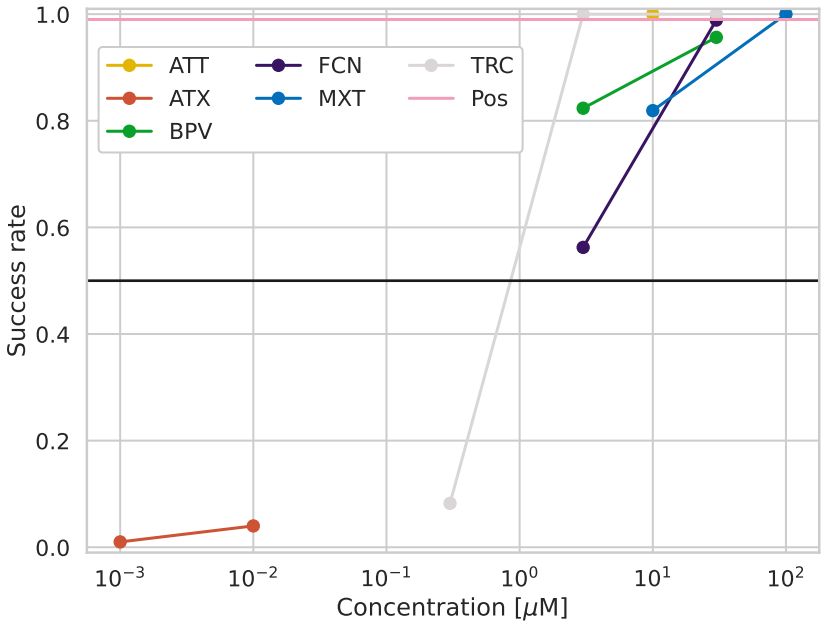

Figure 10.17: Third study case: Classification details for 'Hit' compounds. 'Pos' stands for the positive control: Tetrodotoxin at $1\mu M$.

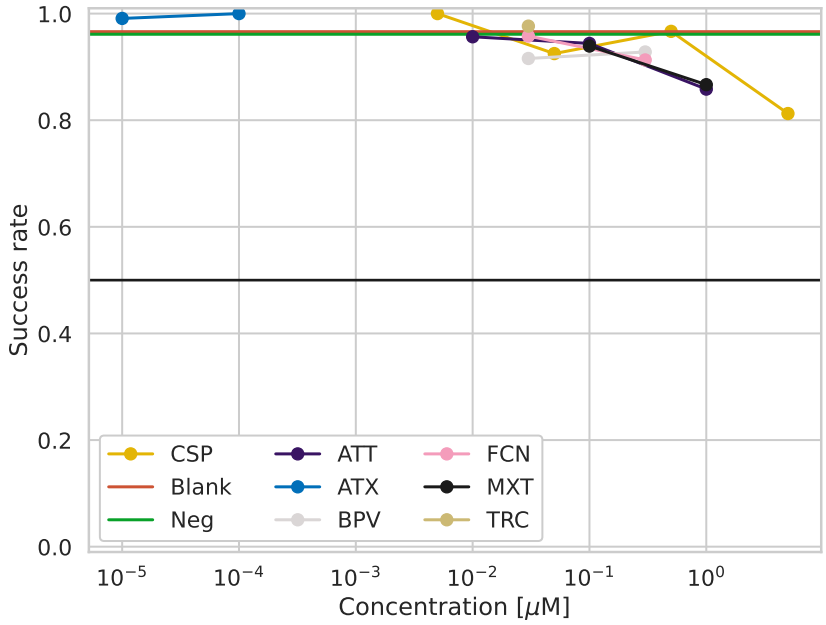

Figure 10.18: Third study case: Classification details for 'No Hit' compounds.

Name	Frequency (%)
Amp	21.4
$Amp * Ch_{50}$	18.9
$Ch_{25} * Ch_{50}$	10.7

Table 10.8: Third study case: Frequencies ($\geq 10\%$) of selected entries over the 100 runs. Ch_X denotes the average electric charge at $X\%$ of the beat period and Amp the average maximal amplitude of the sweep.

Chapter 11
Conclusions

The DGDR method seems to be a suitable strategy for safety pharmacology assessment based on patch-clamp techniques with an easy set up. For studies performed in this part, the following steps are always the same:

- Record signals at baseline and under compound addition.
- Extract n_g quantities from collected signals.
- Construct a dictionary considering control and compound scenarios.
- Define a question to answer and label known data.
- Run the DGDR method.

Its extension to regression tasks by changing the score function by an ℓ^2 norm allows the estimation of ionic channel activities from *in silico* AP signals. In this case, labels are the ionic channel activity. These estimations considered into a UKF model improve its convergence in terms of speed and accuracy. Moreover, the consideration of a stochastic process and compound simulation inside the *in silico* model preserve the quality of the results.

DGDR method was also validated on *in vitro* data provided by automated patch-clamp signals for a Hit/No Hit classification problem.

Both studies, confirm that DGRD method is well adapted to particularly study sodium channels using whole-cell patch-clamp configuration.

Part IV

Microelectrode arrays studies

Chapter 12
Introduction

The CiPA initiative aims at replacing the 2005 regulatory strategy recommended by the International Conference of Harmonisation guidelines by a safety pharmacological screening combining *in vitro* assay and *in silico* knowledge [CH14, CGB+16, CFG+16, FHAG+16]. In this context, the MEA (see Section 2.3.2) is a good candidate for drug screening. While several studies have already been made on pro-arrhythmia risk monitoring [GLG+15, MDS+18], a very few was performed on ion channel blockade. Moreover, pro-arrhythmic risk classification studies based on MEA signals consider few quantities extracted from field potentials (e.g. field potential duration). However, there is no guarantee yet that this quantity is the best to answer the raised question.

This chapter aims at classifying MEA signals in a safety pharmacological context described above:

- Arrhythmia risk.
- Ion channel blockade.

The first chapter of this part is the result of a collaboration with Tessa De Korte and Stefan Braam members of Ncardia[1] company, leading to a publication in PLOS Computational Biology [RDKL+20]. This work was the precursor of the methodology developed at the beginning of this manuscript (see Section 5).

The second chapter of this part is devoted to the coupling of the two methodological works. A first application on the Ncardia dataset in the context of potassium channel blockade classification was performed in order to validate the proposed strategy. A second application resulting in a collaboration with Udo Kraushaar from NMI was performed on a larger dataset based on the 28 reference compounds defined by the CiPA.

[1] Leiden, Netherlands. ncardia.com.

Chapter 13

Oriented dimension reduction method to assess ion channel blocking and arrhythmia risk in hIPSC-CMs

Novel studies conducting cardiac safety assessment using human-induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs) are promising but might be limited by their specificity and predictivity. It is often challenging to correctly classify ion channel blockers or to sufficiently predict the risk for Torsade de Pointes (TdP). In this study, we developed a method combining *in vitro* and *in silico* experiments to improve machine learning approaches in delivering fast and reliable prediction of drug-induced ion-channel blockade and pro-arrhythmic behaviour. The algorithm is based on the construction of a dictionary and a greedy optimisation, leading to the definition of optimal classifiers. Finally, we present a numerical tool that can accurately predict compound-induced pro-arrhythmic risk and involvement of sodium, calcium and potassium channels, based on hiPSC-CM field potential data.

Contents

13.1 Introduction

The Comprehensive in vitro Proarrhythmia Assay (CiPA) is an initiative for a new paradigm in safety pharmacology to redefine the non-clinical evaluation of Torsade de Pointes (TdP) [CJVJS16, MDS+18, YKI+18].

It aims to more precisely assess TdP risk *in vitro* by using a multifaceted approach that combines *in vitro* evaluations of electrophysiological responses in human-induced pluripotent stem cell-derived cardiomyocytes (hiPSC-CMs) and *in silico* models providing reconstructions of drug effects on ventricular electrical activity [DCB+17, PHdK+18].

Since CiPA, *in vitro* studies using hiPSC-CMs become an increasingly integrated part of today's cardiac safety assessment. While encouraging, adequately predicting TdP risk of unknown drugs based on *in vitro* studies alone is challenging [BDM+18]. Besides, the analysis of the large data sets derived from those studies is often far from being automated.

One of the main challenges in proposing a high-throughput screening based on novel devices is often related to the variability of the signals measured, that could pose sensible questions about the ability to extract useful information from them. The main impact of the present work is related to this aspect, and the proposed framework can be considered as a first preliminary step towards the setup of a systematic procedure.

The main focus of the present study is to investigate a computational tool that combines statistical analysis and machine learning approaches (used in this context in [LS16]) to the mathematical modelling and the numerical simulations (*in silico* experiments) of the drug effects on the field potential (FP) of hiPSC-CMs obtained by multi-electrode array (MEA) technology.

Two problems of interest in the safety pharmacology community will be addressed: the first one is related to the prediction of the pro-arrhythmic behaviour of a drug, and the second one to the ion channels blockade. These are typical classification tasks. Some classification studies in cardiac electrophysiology were proposed in the literature, on simulated action potentials [LS16, PBL+17] or ECG [BW93, ASM08].

The contributions of the present paper are the following:

1. A dictionary based greedy optimisation method is proposed, that selects the most pertinent signal features to maximise the classification score. This procedure helps correct the classical markers used to analyse Field Potential signals and provides encouraging results.

2. The *in vitro* dataset is complemented by an *in silico* dataset. This makes it possible to explore all the possible scenarios and help mitigate the high-dimensional/low sample size regime potentially affecting the performances of the classifiers.

3. In constructing the signal database, the uncertainties affecting the experimental setup are accounted for. Despite many variability sources, the proposed approach aims at defining a robust classifier. Concerning the problems considered in the present manuscript, there is enough information in the Field Potential to provide

an answer to them, irrespective of all the uncertainties affecting the experimental setup.

4. The proposed approach was tested on real data coming from actual experiments performed with MEA technology.

The work is structured as follows: the first part is dedicated to the methods used to reproduce *in silico* physiological signals (FP and calcium transient signals) based on the bidomain equations [Tun78b] and the O'Hara, Virág, Varró and Rudy (ORd) ionic model [OVVR11]. The relation between drug concentration and ion channel activity is rendered through scaling factors depending on $IC50$ values (as proposed in [MCS+11, ZBS+13, BPS+06]. The outputs of the *in silico* model are the simulations of the Field Potentials (FP) recorded from extracellular micro electrodes, and the averaged calcium transient on a well ($[Ca^{2+}]_i$).

The second part is dedicated to the description of the method used to integrate *in silico* experiments and *in vitro* data in order to design an optimised classification tool. The proposed approach is based on the construction of a dictionary of linear and non-linear forms applied to the set of *in vitro* and *in silico* data; a greedy algorithm is defined to build a sparse observation-to-prediction relation.

Finally, we applied the classification process in two situations: detecting torsadogenicity (TdP risk versus non-TdP risk) with a synthetic dataset and detecting ion channel blockade (for sodium, calcium or potassium channels) by the action of a given compound, on *in vitro* MEA data.

The classification results obtained show that the double greedy optimisation strategy is effective in improving classifiers performances (with only a few parameters to be tuned) and is well adapted to study compound effects on hiPSC-CM electrophysiology that will aid in early and predictive cardiac safety assessment.

13.2 Material & Method

In this section, we present the method developed to improve the classification of electrophysiological regimes based on MEA signals. It consists in fusing together information coming from available experimental MEA data and numerical simulations in order to design the classifiers to be used.

First, the experimental methods are described; then, we show the different models used to reproduce FP and calcium signals (see Section 13.2.2) and we end the section by presenting the optimised classification algorithm (see Section 13.2.4) and the definition of the dictionary entries (see Section 13.2.3). The structure of this section is shown in Figure 13.1.

13.2.1 Experimental setup

The methods used to perform the experiments and acquire the recordings of the FP are presented in detail below.

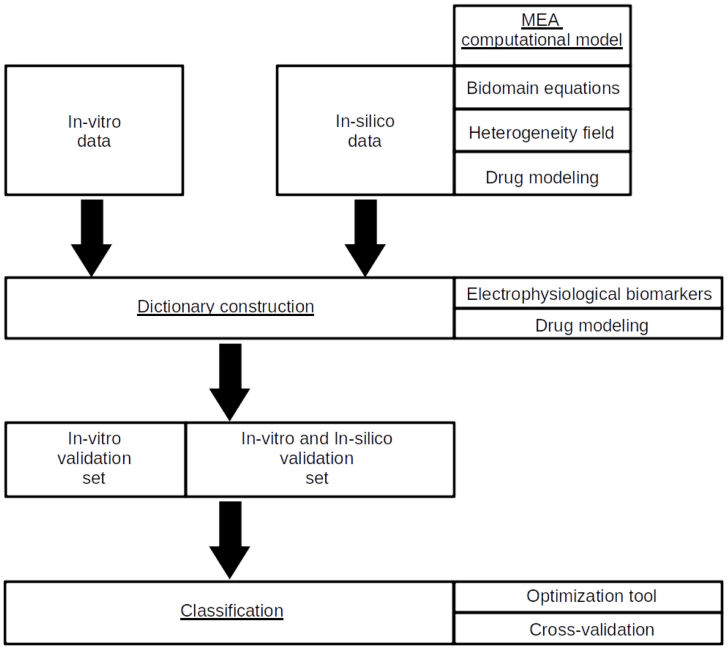

Figure 13.1: Scheme of the Materials and Methods section.

13.2.1.1 Cell culture

Human iPSC-CMs (Pluricyte Cardiomyocytes, Ncardia, Leiden, The Netherlands) were stored in liquid nitrogen until thawed and cultured onto 96 well MEA plates (Axion Biosystems, Inc., Atlanta, USA) according to manufacturer instructions (see Section 2.2.3 for more details on hiPSC). Briefly, the MEA plates were coated with fibronectin ($50\mu g/mL$ in PBS [+Ca2 + & Mg2+], Sigma-Aldrich, St. Louis, MO, USA; Cat. No. F-1141) for 3 hours at $37°C$ and 5% CO2. After 3 hours of incubation time, the excess of fibronectin coating solution was removed and cells were plated in a $5\mu L$ droplet at a density of 25000 cells per well. After 1 hour of incubation ($37°C$ and 5% CO2), $100\mu L$ pre-warmed ($37°C$) medium (Pluricyte Cardiomyocyte Medium, Ncardia, Leiden, The Netherlands) was carefully added to each well. Cells were maintained in Pluricyte Cardiomyocyte Medium for 8 days and refreshments took place at day 1 post-thaw and subsequently every other day. MEA recordings were performed at day 8 post-thaw. The choice of these different parameters of the experimental setup was presented and commented in [ZdKN+19].

13.2.1.2 Test compounds

At day 8 post-thaw, medium was refreshed at least 2 hours before compound addition. The 12 test compounds were provided by the Chemotherapeutic Agents Repository of the National Cancer Institute and consisted of a random subset of CiPA compounds. The compounds were from 3 clinical TdP risk categories: low/no (Loratadine, Mexile-

Compound*	$IC50$ (μM)**			Concentration (μM)				Tr/V	Label
	hERG	Cav1.2	Nav1.5	#1	#2	#3	#4		
Azimilide	$< 1^{\dagger}$	$17.8^{\dagger}$	$19^{\dagger}$	0.01	0.1	1	10	V	K [BEJ+98], Na [YT97], Ca [YT97]
Bepridil	0.16	1.0	2.3	0.01	0.1	1	10	V	K [HMEHhZ08], Ca [YBS86], Na [YBS86]
Chlorpromazine	1.5	3.4	3.0	0.0951	0.3004	0.9494	3	V	K [HMEHhZ08], Ca [CVF84], Na [ONN89]
Cisapride	0.02	11.8	337	0.0032	0.01	0.0316	0.1	V	K [HMEHhZ08]
Clarithromycine	32.9	> 30	NA	0.1	1	10	100	V	K [HMEHhZ08]
Clozapine	2.3	3.6	15.1	0.0951	0.3004	0.9494	3	Tr	K [LKK+06], Ca [NHW+17]
Diltiazem	13.2	0.76	22.4	0.01	0.1	1.0	10	Tr	Ca [LT83]
Dofetilide	0.03	26.7	162.1	0.0003	0.001	0.0032	0.01	V	K [HMEHhZ08]
Droperidol	0.06	7.6	22.7	0.03169	0.10014	0.31646	1.0	Tr	K [HMEHhZ08]
Ibutilide	0.018	62.5	42.5	0.0001	0.001	0.01	0.1	Tr	K [HMEHhZ08]
Loratadine	6.1	11.4	28.9	0.001	0.003	0.0095	0.03	V	K [Cru00], Ca [NHW+17]
Mexiletine	62.2	125	38	0.1	1.0	10	100	Tr	Na [MWZ+13], K [GTBR+15]

* Colours corresponds to the TdP risk (green: low, orange: medium and red: high). See Colatsky *et al* [CFG+16].
** From Ando *et al* [AYY+17]. † From Yao *et al* [YT97].
Tr/V: Considered into the Training (Tr) or Validation (V) set.

Table 13.1: Experimental data information.

tine, Diltiazem), intermediate (Clozapine, Chlorpromazine, Clarithromycine, Cisapride, Droperidol) and high (Ibutilide, Dofetilide, Bepridil, Azimilide) [CFG+16] (see Table 1). Chemical stock solutions at 1000-fold of the target concentrations were prepared under sterile conditions in DMSO and stored at $-20°C$, according to HESI Myocyte Phase II Validation Study Protocol instructions. The serial diluted compounds were further prepared in DMSO on the day of compound assay. The 10-fold final dilutions of the compounds were prepared with Pluricyte Cardiomyocyte Medium, for single time use only. Pluricyte Cardiomyocytes were exposed to four different concentrations of the compound, under sterile conditions in single point additions (i.e. one concentration per well) in five replicates for each concentration. Vehicle control was 0.1% DMSO. All the experiments were performed under permits granted from the Commissie Medische Ethiek Leiden University Medical Center (permit number: NL45478.058.13).

13.2.1.3 MEA recordings

At day 8 post-thaw, 96 well MEA plates seeded with hiPSC-CMs were placed in the Maestro MEA device (768-channel amplifier) with an integrated heating system, temperature controller and data acquisition interface (Axion BioSystems, Inc., Atlanta, USA). The field potential traces of the hiPSC-CMs were recorded prior to (baseline) and $30min$ after compound addition for $5min$. The recording conditions were at 37°C using Cardiac Standard filters and amplifiers in spontaneous cardiac mode ($12.5Hz$ sampling frequency, $2kHz$ Kaiser Window, $0.1Hz$ IIR). The beat detection threshold was $300\mu V$.

13.2.2 MEA computational model

This part provides a detailed description of the mathematical models used to simulate FP in a realistic MEA geometry. Simulated FP studies were already performed for *in silico* assessment of drug effects [ZBS+13] or channel activity identification [RBZ+17] and have shown the potency to reproduce and analyse compound effects on cardiac electrophysiology.

The first section concerns the bidomain equations, which governs the electrical activity propagation in a tissue. Since the cells might not be perfectly uniformly distributed in a well and the cell population might even be heterogeneous, a stochastic model of the population distribution was adopted, which is described in Section 13.2.2.1. In the last part (13.2.2.2) we describe the compound simulation strategy, aiming at reproducing the experimental protocol used to classify reference compounds holding ion channel blocking properties.

To simulate MEA recordings, we consider the same model as the one presented in Section 3.3 with the following parameters and methods:

- Bidomain equations (see Equation (3.7)):
  - The discretisation of the partial differential equation was done in space using P1 Lagrange finite elements and in time using backward differentiation formula (BDF) schemes with a time step of $0.1ms$.
  - The ODE part governing the action potential modelling (I_{ion}) was solved using PDF scheme with adaptive time steps and order, whose implementation is provided by Sundials' CVODE[HBG+05]. These space and time discretisations of the bidomain equations were already used in different studies (*in silico* ECG and *in silico* field potentials) and have shown qualitatively good results compared with real data [ABC+18, CCG13, RBZ+17, SCG16, TSC+18, ZBS+13]. The ionic current $I_{\text{ion}}(V_m,\gamma)$ and the state variable γ are provided by the ORd model [OVVR11]. Three types of cells are considered to mimic the monolayer heterogeneity (see Section 13.2.2.1): Epicardial, Mid-myocardial and Endocardial. These cell types are simulated through specific sets of parameters given in [OVVR11]. This model takes into account the main concentration dynamics ($[\text{Na}^+]_\text{i}$, $[Ca^{2+}]_i$ and $[\text{K}^+]_\text{i}$).
  - Parameters of the source term I_{app} model described in Section 3.3.2.2 are $r = 50\mu m$ (the radius of the source), $I_0 = -130pA/pF$ (the maximum stimulation value) and $\sigma = \frac{\Delta t}{6}$ with $\Delta t = 4$ms.
- Electrode model: parameters are given in Table 13.2.

C_{el}	R_i	R_{el}
$1nF$	$2M\Omega$	$10M\Omega$

Table 13.2: Parameters used for the imperfect electrode model.

- Heterogeneity field: see Section 13.2.2.1.
- Drug modelling: see Section 13.2.2.2.

To mimic experimental measurements, a $10\mu V$ standard deviation noise of a zero-mean Gaussian was added to FP. As some devices are able to get the intracellular calcium

transient by fluorescence, we made the assumption that we have access to intracellular calcium transient data. We added a zero-mean Gaussian noise of $10^{-3}\mu M$ on the intracellular calcium transient obtained by simulation with the ORd model.

13.2.2.1 Heterogeneity

The hiPSC-CMs used in this study are $> 70\%$ pure cardiomyocytes based on positive Troponin T (TnT) expression. At least 70% of the TnT positive cells express a ventricular phenotype (based on ventricular myosin light chain 2 (MLC2v) expression and patch clamp technology). The other 30% of the cell population is of mesodermal origin. The actual distribution of these cells inside the well is unknown, and is a source of uncertainty that we need to take into account when developing the classifier in order to provide meaningful results in realistic applications.

Here, we consider the strategy developed in Section 3.3.4. When discretised on the finite element space (P1 Lagrangian elements were used), a cell type was affected to each node of the finite element mesh according to the following rule:

$$\text{CellType}_\text{i} = \begin{cases} \text{Epicardial, if } c_i < \frac{1}{3} \\ \text{Mid-myocardial, if } c_i > \frac{2}{3} \\ \text{Endocardial, otherwise} \end{cases}, \tag{13.1}$$

where $c_i \in [0,1]$ is given by the random process c discretised at the node whose coordinates are $\mathbf{x}^{(i)}$ (see Equations (3.11) and 3.12). Example of random heterogeneity obtained with this method is presented in Figure 13.2.

A photo of the well corresponding to the right panel of Figure 13.2 is shown in Figure 13.3.

13.2.2.2 Drug modelling

In this study we assume that a drug may affect only sodium, calcium and/or potassium channels. The conductance-block model is rewritten in Equation (13.2). We refer to Section 3.2.2 for more details.

$$g_s = g_{control,s}\left[1 + \left(\frac{[D]}{IC50_s}\right)^{n_H}\right]^{-1}, \tag{13.2}$$

meaning that, for this study, $s \in \{Na, Ca, K\}$.

Here, we chose to set the Hill coefficient n_H at 1. The first reason is due to the confidence intervals of computed Hill coefficients for different compounds [CDM+17] which most of the time includes 1. The second reason comes from the use of the EFTPC in our simulations. Varying the Hill coefficient between 0.6 to 1.4, the standard deviation of the channel activity is lower than 0.05 for concentrations higher than the $IC50$ and lower than 0.03 for concentrations lower than the $IC50$ (see Figure 13.4). The use of the EFTPC leads to a low variability in the channel activity according to the Hill coefficient and studied compounds.

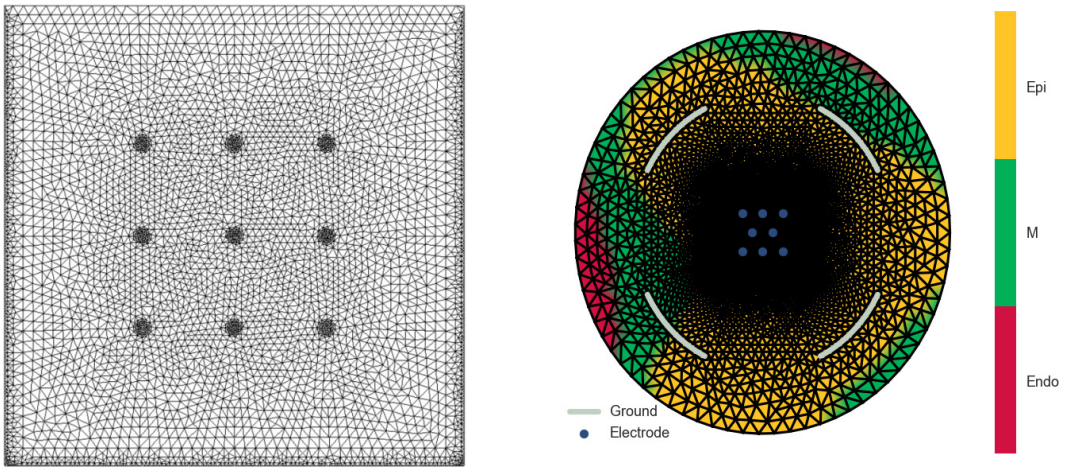

Figure 13.2: Finite element meshes of MEA used and example of heterogeneity field. Left: Finite element mesh representing one well including 9 electrodes of the 6-well MEA device from Multichannel Systems (used in Section 13.3.1). MEA device documentation is available on: http://www.qichi-instruments.com/bookpic/20163120452599.pdf. Right: Finite element mesh representing one well including 8 electrodes of the 96-well MEA device from Axion Biosystems with an example of generated cell heterogeneity field (used in Section 13.3.2). MEA device documentation is available on: https://www.axionbiosystems.com/sites/default/files/resources/mea_plates-brochure-rev_06.pdf.

13.2.3 Dictionary entry computations

The details about the construction of the dictionary entries are provided and commented in this section. As mentioned previously, the dictionary is a collection of linear and non-linear forms applied to the signals, corresponding to the definition of features (think, for instance, to the maximum of the signal, or its average, and so on).

The greedy optimisation strategy has been devised to project into a as low as possible dimensional subspace with a sparse contribution of the entries. This internal stage if the descent is easily parallelisable, which allows an affordable dictionary size potentially large (a few hundred in the study, potentially few thousands) in order to avoid a possible loss of information.

In the present work, the dictionary is divided into two parts:

- non-agnostic, or informed.
- agnostic.

In the informed part, we collect the biomarkers extracted from the signal, identified by the experts as correlated to some regime of interest. These quantities are meant to reveal a particular state of the system or alteration of a parameter. For instance,

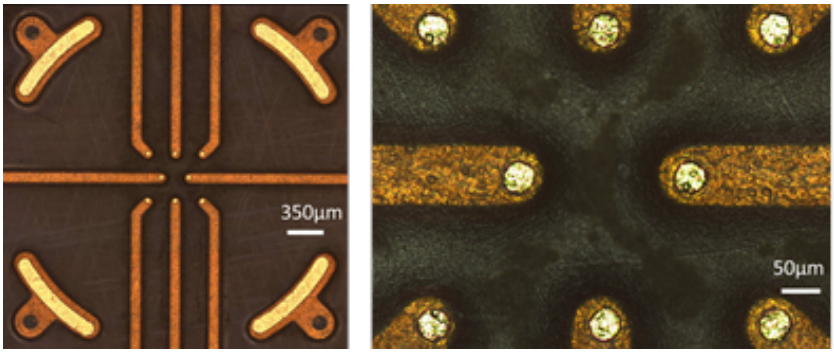

Figure 13.3: Photo of one well corresponding. The corresponding P1 Lagrange finite element mesh is shown in the right panel of Figure 13.2.

altering the sodium channel activity induces a modification in the depolarisation amplitude [TBVM+18]. The second part of the dictionary is agnostic, meaning that the linear and non-linear forms introduced are extracted from the signal as a mathematical object. The goal of the agnostic part of the dictionary is to enrich it, henceforth increasing the possibilities of computing from the dictionary an input leading to a good classification. The dictionary entries and their numbering are presented in Table 13.5.

13.2.3.1 Electrophysiological biomarkers

First, some intuitive biomarkers were extracted, e.g. depolarisation amplitude (DA), field potential duration (FPD), etc. These quantities (called parameters in the electrophysiology community are presented in Figure 13.5.

Remark 20
In the electrophysiology community we often refer to parameters to designate quantities extracted from the experimental signals. In the present work, we follow the usage in applied mathematics and engineering communities, that refers to "parameters" the quantities affecting the state of the system and not the quantities read from the observable system.

Their computation follows the work in [TRLG18] and it is described in more details in Section 13.5.1. Concerning the calcium transient signal computation, details are given in Section 13.5.2.

As these values of these biomarkers are computed in control and drug case, we decided to use relative values to the control case. For instance, the DA ratio is: $\frac{DA_{drug}}{DA_{ctrl}}$. The justification of this choice is shown in Figure 13.6. As we can see, even if the control case is different, the impact due to a compound is qualitatively the same regardless of the heterogeneity field.

An example of an effect of a drug on the repolarisation of the cells compared to baseline is presented in Figure 13.7. In a case where a drug does not affect the repolarisation, we

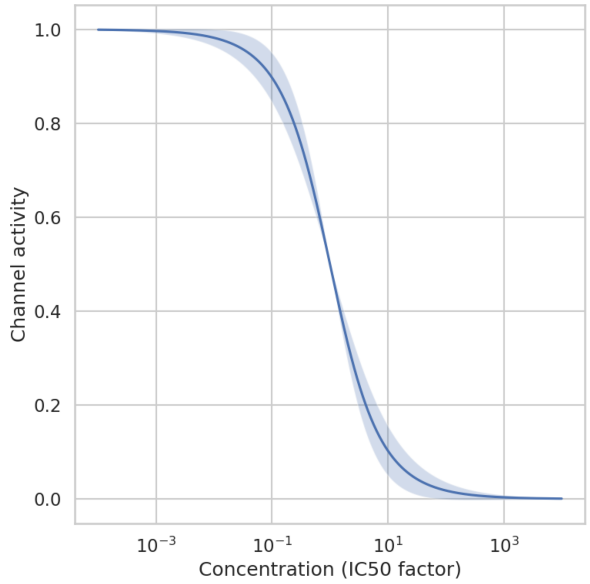

Figure 13.4: Channel activity average and standard deviation for a Hill coefficient varying from 0.6 to 1.4. The abscisse is the concentration factor with respect to the $IC50$.

should obtain a curve similar to $f(x) = x$ (red line in Figure 13.7).

In the case where the repolarisation is affected by a compound, a distortion appears on the signal (see black dots lower panel in Figure 13.7 corresponding to an increase in the FPD). Five markers (from K1 to K5) were extracted from the signal, with FP_{rep}^{drug} and FP_{rep}^{ctrl} the repolarisation part of the FP for the drug and control case:

- Maximum distance: $max_i \sqrt{\left(FP_{rep}^{drug}(i) - FP_{rep}^{ctrl}(i)\right)^2}$.
- ℓ^2 norm: $\left|\left|FP_{rep}^{drug} - FP_{rep}^{ctrl}\right|\right|_{\ell^2}$.
- Average deviation: $\frac{1}{N}\sum_{i=1}^{N}\left(FP_{rep}^{drug}(i) - FP_{rep}^{ctrl}(i)\right)$.
- Maximum deviation: $max_i\left(FP_{rep}^{drug}(i) - FP_{rep}^{ctrl}(i)\right)$.
- Time of the maximum deviation.

13.2.3.2 Wavelet coefficients

In order to construct the agnostic part of the signal, a wavelet decomposition was considered for the repolarisation phase. The number of coefficients retained is such that the signal could be represented up to the noise level by the wavelets expansion.

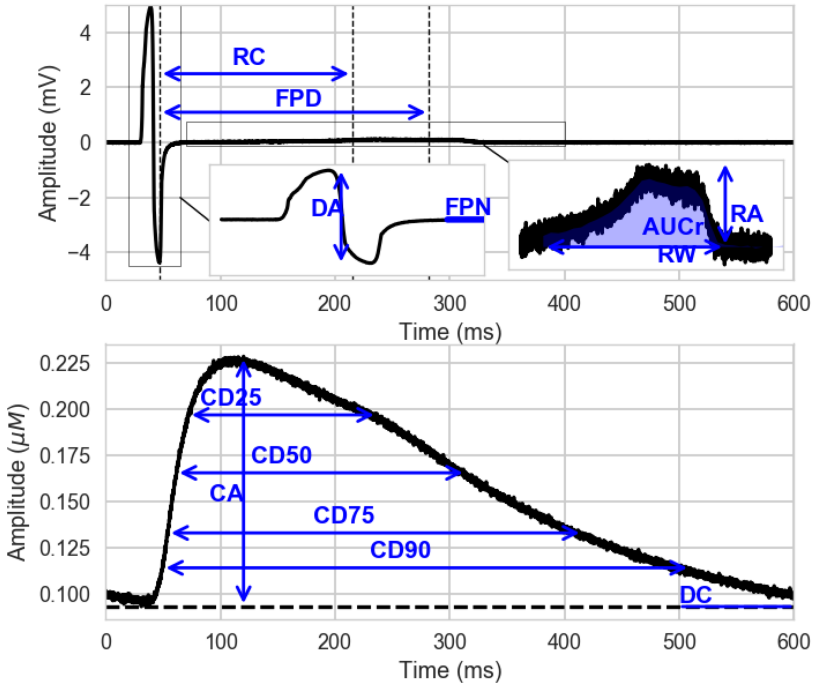

Figure 13.5: List of the parameters computed on FP (up) and Calcium transient (down). RC: Repolarization Centre; FPD: Field Potential Duration; DA: Depolarisation Amplitude; FPN: Field Potential Notch; AUCr: Area Under Curve of the repolarisation wave; RA: Repolarization Amplitude; RW: Repolarization Width; CA: Calcium Amplitude; DC: 'Drowsing Calcium'; CDX: Calcium Duration

When a new signal is analysed, only the selected coefficients (already computed for the training database) are then used to reconstruct the signal. If the L^2 error is lower than an arbitrary value, we store these coefficients. Otherwise, we compute the new location and add the missing locations. The wavelet transform was done on the absolute difference between the drug case and the control case by considering Daubechies level 8 wavelets. Despite this choice leads to a Gibbs phenomenon, it induces a smoother reconstructed signal if compared with Haar wavelets [BL13]. An example of reconstruction is shown in Figure 13.8. The algorithm to get the positions is presented in the pseudo-code 7.

13.2.4 Classification

Given a molecule which is a candidate to become a drug, several questions arise concerning its impact on the electrical activity of cells. Basic questions like: *Is this drug blocking channel X?* or: *Is the drug potentially causing arrhythmia?* are naturally treated by solving a classification problem.

One of the main difficulties related to such a study is the curse of dimensionality [Bel15] since we are dealing with high dimension, low sample size data. Otherwise stated, the

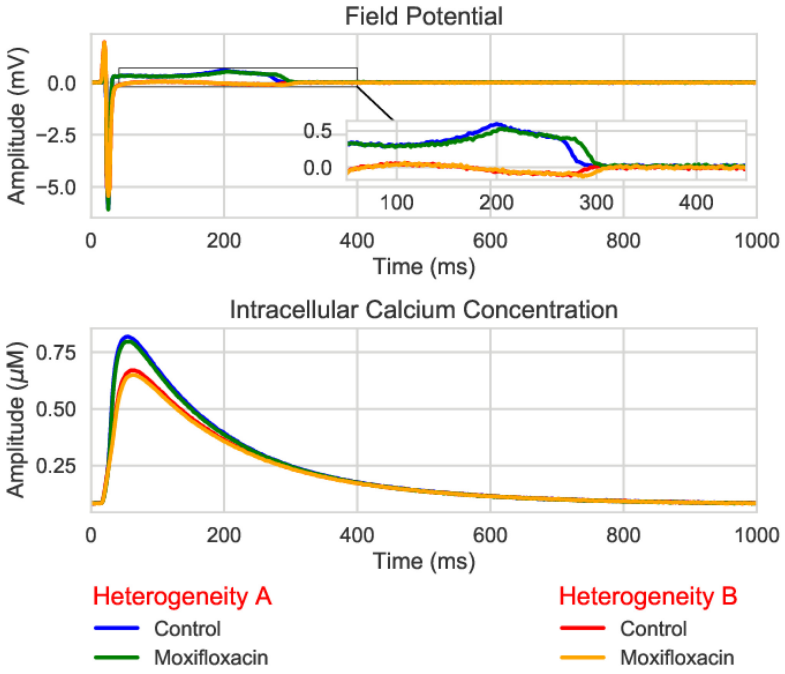

Figure 13.6: Moxifloxacin simulation. Simulation of the effect of Moxifloxacin at effective free therapeutic plasma concentration ($10.96\mu M$, see Table 13.6) on the FP (from one electrode) and intracellular calcium transient (from one well) for two different heterogeneity fields. A finite element mesh of 96-well MEA device from Axion Biosystems was used for this simulation (see right panel of Figure 13.2).

function to be identified in view of setting up efficient classifiers is defined over a high dimensional domain and the number of available data is too low. To tackle this problem, numerical simulations were exploited. The rationale behind the strategy is twofold: first, we added virtual *in silico* experiments to the data set to increase the population size. Second, we exploited the simulations to extract meaningful low dimensional subsets of the data, which contributed to the mitigation of the high-dimensionality. Several methods are available in the literature to extract these low dimensional subsets [SIL07]. However, the risk in using generic problem independent methods is that the subsets obtained could drastically reduce the amount of information conveyed about the quantity of interest to be classified. Henceforth, we constructed a low dimensional subset of the data that has been exploited in the classifier construction, designed to deliver optimal classification performances. This method can be applied to all different classification techniques, and the result is classifier dependent.

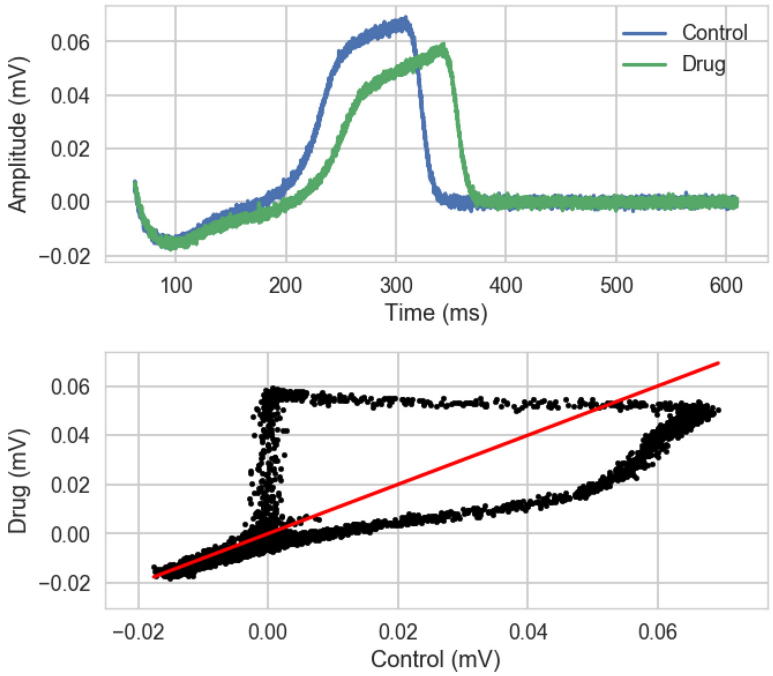

Figure 13.7: Extended dictionary based on repolarisation. Upper panel: FP repolarisation. Lower panel: Repolarisation of cells affected by a compound with respect to the control case repolarisation. The red line corresponds to the case where the repolarisation is not affected.

13.2.4.1 Classification optimisation

The goal-oriented dimension reduction method used in this study was a precursor of the DGDR method described in Section 5. Let n_s be the sample size, $y_*^{(i)} \in \{-1,1\}$ the true label of the i^{th} sample and $\widehat{y}^{(i)} \in \{-1,1\}$ the predicted label of the i^{th} sample associated with $\widehat{p}^{(i)} \in [\frac{1}{2},1]$ the confidence of the classifier to be the predicted label. Let n_1 and n_{-1} be the number of samples labelled 1 and -1 respectively, introduced to avoid a possible bias due to unbalanced classes. Then, we have $n_s = n_1 + n_{-1}$. Let $\delta_i(x)$ be the Dirac function ($\delta_i(c) = 1$ if $c = i$, 0 otherwise). Finally, parameter $\alpha \geq 1$ is introduced to penalise the false positive case (i.e. $y_*^{(i)} = 1$ and $\widehat{y}^{(i)} = -1$). The expression of the cost function is presented in Equation (13.3).

$$\begin{aligned}\mu = -\frac{1}{n_s}\sum_{i=1}^{n_s}\widehat{p}^{(i)}\Big[&\frac{n_s}{n_1}\delta_1(\widehat{y}^{(i)})\delta_1(y_*^{(i)}) + \frac{n_s}{n_{-1}}\delta_{-1}(\widehat{y}^{(i)})\delta_{-1}(y_*^{(i)}) \\ &- \alpha\frac{n_s}{n_1}\delta_{-1}(\widehat{y}^{(i)})\delta_1(y_*^{(i)}) - \frac{n_s}{n_{-1}}\delta_1(\widehat{y}^{(i)})\delta_{-1}(y_*^{(i)})\Big].\end{aligned} \tag{13.3}$$

This cost function can easily be bounded (see Proposition 13.1) as shown in the

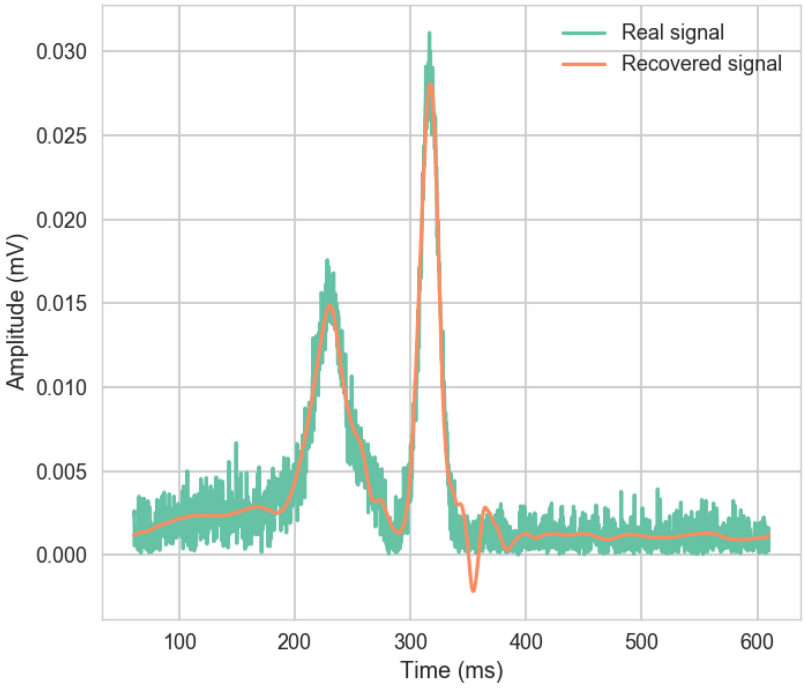

Figure 13.8: Reconstruction of the absolute difference between the drug and control signals for the plateau and repolarisation phases, based on wavelets coefficients.

Section 13.5.2 in the Appendix.

Proposition 13.1
Given the cost function defined in Equation (13.3)*, we have:* $\mu \in [-2, 1+\alpha]$.

The rationale of including the terms $\hat{p}_i$ in the cost function is to better describe the performances of the classifier, accounting for the confidence in the classification, and not merely on the success rate. This aims at setting up a robust classification tool. Following the same principles, the cost function can be extended to c classes as shown in Equation (13.4):

$$\mu = -\frac{1}{n_s}\sum_{i=1}^{n_s}\widehat{p}^{(i)}\left\{\sum_{j=1}^{c}\left[\frac{n_s}{n_j}\delta_j(\widehat{y}^{(i)})\delta_j(y_*^{(i)}) - \frac{n_s}{n_j}\alpha_j\sum_{\substack{m=1\\m\neq j}}^{c}\delta_m(\widehat{y}^{(i)})\delta_j(y_*^{(i)})\right]\right\}, \tag{13.4}$$

where n_j is the number of samples labelled j and $\alpha_j > 0$ the weight assigned if the predicted label is not the class j. The bounds of μ in this general case are given in Proposition 13.2 and its demonstration in Section 13.5.2 in the Appendix.

Proposition 13.2
Given the cost function defined in Equation (13.4)*, we have:* $\mu \in [-c, \sum_{j=1}^{c}\alpha_j]$.

Algorithm 7 Wavelet coefficient.

n_s {Number of signals to compute positions.}
thr {Threshold for the wavelets transform.}
v_p {Empty array of positions.}
for $i := 1$ **to** n_s **do**
  f_i {Get the i^{th} signal.}
  $c_{wvlt} = CWT(f_i, thr)$ {Computes wavelets coefficients.}
  v_p^{nz} {Get the non-zeros positions of c_{wvlt}.}
  if i==1 **then**
    $v_p = v_p^{nz}$
  else
    $v_p = v_p \bigcup v_p^{nz}$
  end if
end for

In the case where we do not penalise classes, all the α_j are equal to 1. A regularisation term was added to the cost function:

$$\mu_{reg} = \mu + \beta \left(\sum_{j=1}^{k} (1 - \sum_{i=1}^{N_b} \omega_{ji}^2) \right), \tag{13.5}$$

where $\beta \in \mathbb{R}^+$ is a penalisation parameter and k the dimension of the projected subspace. This term aims at breaking the scaling invariance of the linear combination of the dictionary entries. In particular, if a linear classifier is used, let $\alpha \in \mathbb{R},\ \alpha \neq 0$, the classification score when using $\alpha\Omega$ is the same as Ω, irrespective of the value of α. The optimisation problem reads:

$$\Omega^* = \arg \inf_{\Omega \in \mathbb{R}^{d \times n_g}} \mu_{reg}.$$

The optimisation problem is challenging to be solved for two main reasons. First, the dimension k of the input space to be found is not known *a priori*, but has to be determined. Second, given realistic signals, the number of features that can be extracted, and hence the size of the dictionary, can be quite large, leading to an optimisation problem on a large-dimensional space. To mitigate these difficulties, a greedy optimisation strategy was adopted has proposed in the DGDR method (see Section 5.2.2.2).

13.2.4.2 Cross-validation

As inputs are computed to maximise the classification, a risk is to lose the generalisation capacity of a good classifier. To prevent the overfitting and to increase the robustness of the strategy a random K-fold cross-validation was used. A stratification was applied on the data to ensure the conservation of the output repartition in each fold.

The pseudo-code is described in Algorithm 8 and the corresponding Scikit-Learn method was used.

Algorithm 8 Randomised K-fold cross-validation procedure.

M: input matrix
l: output vector
$n_{fold} = 2$ {Number of folders for each K-Fold.}
$n_{kfold} = 500$ {Number of K-Fold.}
$E = (1,2,...,n_s)$ {Sample numbering.}
$Cnt = 0$ {Initialise counter vector of size n_s.}
$P = 0$ {Initialise matrix $P \in \mathbb{R}^{n_s \times 2}$ with 0 values.}
for $i := 1$ **to** n_{kfold} **do**
    $E' = getFolders(E,n_{fold},l)$ {Generate the n_{fold} folders with respect to the stratification.}
    for $j := 1$ **to** n_{fold} **do**
        $pos_{test} = E'[j]$ {Testing folder (vector of indices).}
        $pos_{train} = E \backslash pos_{test}$ {Complementary of pos_{test}.}
        $M_{train} = M[pos_{train}, :]$ {Extract train submatrice.}
        $l_{train} = l[pos_{train}]$ {Extract train subvector.}
        $M_{test} = M[pos_{test}, :]$ {Extract test submatrice.}
        $l_{test} = l[pos_{test}]$ {Extract test subvector.}
        $clf =$ Train on (M_{train},l_{train}) {Train the classifier.}
        $proba =$ Test on (clf, M_{test}) {Test the new data with the classifier.}
        $P[pos_{test}, :] = P[pos_{test}, :] + proba$ {Add the new probabilities.}
        $Cnt[pos_{test}] = Cnt[pos_{test}] + 1$
    end for
end for
for $i := 1$ **to** n_s **do**
    $P[i, :] = P[i, :]/Cnt[i]$ {Compute averaged probability for the i^{th} sample.}
end for

The repetition of the random K-fold strategy allows the convergence of the weights regardless of the training and test set generated. The higher the number of weights to determine, the higher should be the number of random K-fold.

13.3 Results

Two different studies were performed in the present work: classify compounds for their risk on TdP; classify compounds for their ion channel blocking properties. The results of these are presented hereafter.

In the first part of Section 13.3.1, we describe the study results based on the conductance-block model (see Equation (3.5)). Using this model, we classified the TdP risk of 86 known compounds based on simulated data using the compound's $IC50$ values

for blocking sodium, potassium and calcium currents and the effective free therapeutic plasma concentration (EFTPC) values, reported by the literature.

In the second study (see Section 13.3.2) we classified compounds based on experimental data. The outcome consists in identify which channel is affected (sodium, calcium or potassium) by a compound. These experiments were performed for 12 compounds using Pluricyte Cardiomyocytes. Five of them were used for the training set and seven of them for the validation. Because of the low sample size of data, a simulated database was generated to enrich the training set.

The stop criterion used for the following results is when the cost variation between the last two components is lower than 5%. The penalisation parameter β described in Equation (13.5) was set to 0.1.

13.3.1 TdP classification

This section is dedicated to the torsadogenicity risk classification. Only simulated data are considered for this study. To predict the risk of TdP of a wide range of compounds, we simulated the application of 86 known compounds previously reported by [LS16].

13.3.1.1 Tests setup

The numerical choices leading to the results are summarised hereafter:

- The false positive part in the cost function (see Equation (13.3)) was taken $\alpha = 2$, to minimise the false positive rate.
- Bidomain equation parameters are summarised in Table 13.3.

A_m	C_m	σ_i	σ_e
$200 cm^{-1}$	$1.0 \mu F cm^{-2}$	$5.0 nS cm^{-2}$	$5.0 nS cm^{-2}$

Table 13.3: Bidomain equation parameters used for Multichannel Systems MEA device.

- Drugs were modelled using Equation (3.5) presented in Section 13.2.2.2. The $IC50$ values for each compound are given in [MCS+11, KOPM+13]. Concentrations chosen to simulate compounds are the effective free therapeutic plasma concentrations (EFTPC). These values are listed in Tables 13.7 and 13.6. Eighteen drugs were modelled twice because of their different $IC50$ and EFTPC observed in the literature (see Tables 13.7 and 13.6).
- The corresponding channels blocked in the ORd model are I_{Na} (g_{Na}), I_{Kr} (g_{Kr}) and calcium channels (I_{CaL}, I_{CaNa} and I_{CaK}) through the PCa variable as previously reported in [LS16].
- A 6-well MEA device (Multichannel Systems) with 9 electrodes per well (60-6wellMEA20030iR-Ti) where the corresponding finite element mesh is presented in

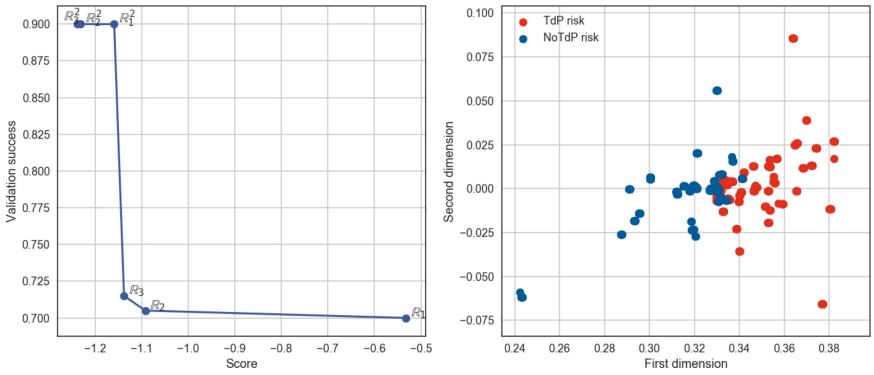

Figure 13.9: TdP risk classification through simulations of 86 compounds. Left: Validation versus Cost curve depending on the number of components and the dimension. Right: Drug repartition in the input space after convergence of the algorithm.

the left panel of Figure 13.2. A cell heterogeneity field was applied on this finite element mesh following the strategy developed in Section 13.2.2.1.

- The sparse optimisation was performed on a dataset of 1520 data points (76 first compounds, each compound simulated 20 times with different heterogeneity and sources). The FP traces corresponding to the last 10 compounds were also simulated 20 times with different heterogeneity and sources, but used for the validation set. The same process was done for the calcium transient signals. The dictionary entries used for this classification problem are summarised in Table 13.5.

13.3.1.2 Results of TdP classification

We start this section by commenting on the results of the classifier as function of the input space constructed by the greedy algorithm. In Figure 13.9 the success rate of the classification for the validation set is plotted as function of the cost presented in Section 13.2.4.1, Equation (13.3). The cost minimised by the proposed algorithm is a pertinent descriptor of the success rate of the classifier. The input space selected by progressively increasing the input space dimension as well as the components per dimension produces a high success rate. The input space corresponding to the case where the input is in $\mathbb{R}^2$ (with three dictionary components per direction) is shown in Figure 13.9, from which we can appreciate that the separation between the classes is satisfactory.

The results of the classification are detailed hereafter. Figure 13.10 shows the confusion matrices for the training set (left, in blue) and for the validation set (right, red). Globally, the results are similar for training and validation (no apparent overfitting phenomena were seen). The type II error (wrongly classifying a compound as non-torsadogenic) is well minimised thanks to the choice to penalise false positives ($\alpha = 2$ in the cost function, Equation (13.3)). In the validation set, no compounds were wrongly classified as non-torsadogenic. Only the Propranolol was misclassified as torsadogenic (see Table 13.7).

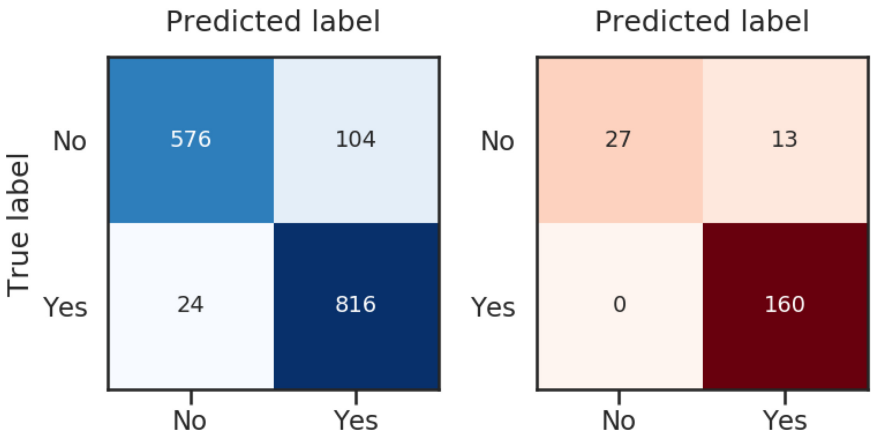

Figure 13.10: Confusion matrices obtained for TdP risk classification of 86 compounds after convergence of the algorithm. Yes: TdP risk. No: No TdP risk. Left: Training set (sample size: 1520) using randomised K-fold cross-validation. Sensitivity = 0.98, Specificity = 0.85 and Accuracy = 0.92. Right: Validation set (sample size: 200). Sensitivity = 1, Specificity = 0.675 and Accuracy = 0.935.

It is interesting to monitor the classification results at different stages of the algorithm. The confusion matrices are given in Figure 13.19. Confusion matrices obtained for test and validation sets show an improved TdP risk classification when we increase the number of components and dimensions. This improvement is particularly visible on the test set for the first components. The training on the first dimension is not sufficient to classify well the validation set, meaning that other dictionary entries would have been selected by the algorithm (for the first dimension) if the validation set was in the training set. However, dictionary entries selected for the second dimension seem to be better to discriminate torsadogenic risk on the validation set.

13.3.2 Channel classification

This section is dedicated to the channel classification of 12 compounds based on *in vitro* data derived from MEA recordings of spontaneous beating hiPSC-CMs (Pluricyte Cardiomyocytes) cultured on 96 well MEA plates (8 electrodes per well, Axion Biosystems), as described in Section 13.2.1.1. As we are limited by the experimental sample size (see compound list in Table 13.1), we enriched the experimental database with a simulated database (for which we know the classification output). For this study only FP traces were recorded and used for the training and classification, no calcium transient measurements were performed.

13.3.2.1 Tests setup

The numerical choices leading to the results are summarised hereafter:

- Bidomain equation parameters are summarised in Table 13.4.

A_m	C_m	σ_i	σ_e
$1200cm^{-1}$	$1.0\mu Fcm^{-2}$	$1.2\mu Scm^{-2}$	$1.2\mu Scm^{-2}$

Table 13.4: Bidomain equation parameters used for Axion MEA device with Pluricyte Cardiomyocytes cell line.

- Drugs were modelled using Equation (3.5) presented in Section 13.2.2.2. The *in silico* database was generated blocking alternatively sodium (g_{Na}), potassium (g_{Kr}) or calcium (PCa) channels of the ORd model at a random percentage between 0% and 50%. Other channels are blocked between 0% to 5% to introduce some variability (e.g. blocking sodium at 35%, calcium at 2% and potassium at 3.5%). An example is shown in Figure 13.11.
- The simulated sample size is 140 (computed from signals resulting from the simulation performed for different heterogeneity fields).
- A 96-well MEA device (Axion Bioystems) with 8 electrodes per well where the corresponding finite element mesh is presented in the right panel of Figure 13.2. A cell heterogeneity field was applied on this finite element mesh following the strategy developed in Section 13.2.2.1.

The experimental data leading to the results are summarised hereafter:

- *In vitro* data used for this part are FP traces recorded from a hiPSC-CM monolayer (Pluricyte Cardiomyocytes, Ncardia) plated on a 96 well MEA plate (8 electrodes per well) Axion Biosystems (Classic MEA 96 M768-KAP-96[1]).
- The 12 "CiPA" compounds listed in Table 13.1 were tested on Pluricyte Cardiomyocytes and FP traces were recorded before and 30 minutes post compound addition. MEA results of 5 compounds were used for the training and MEA results of 7 "blind" compounds for the validation.
- Each compound was tested at 4 concentrations, 1 concentration per well and in 5 replicates ($n = 5$ per concentration).

Using the conductance-block model described in Equation (3.5) we obtain the percentage of activity for each channel and concentration. This is shown in Table 13.8.

Two different kinds of classification problems have been studied. A binary classification (i.e. given a channel, is the molecule affecting its functioning), whose results are shown in Section 13.3.2.2 and a ternary classification (i.e. is the molecule affecting potassium, calcium or sodium?), whose results are reported in Section 13.3.2.3. For the numerical experiments proposed, the success rate of the classifier for the training set was about 90%. In the following, we present in details the results on the *in vitro* data in the validation set.

[1]Documentation available here.

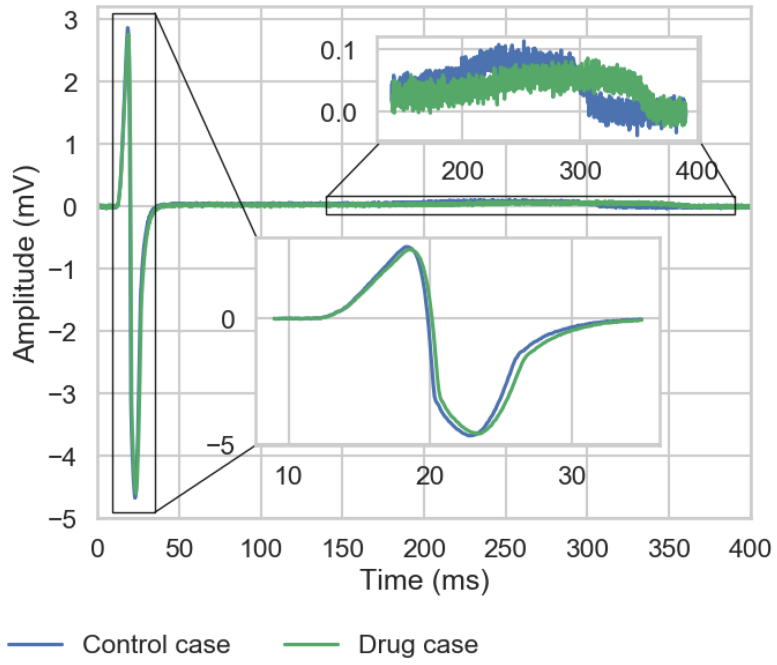

Figure 13.11: Simulated FP under control and compound conditions. FP trace from one electrode, showing the effect of drug simulation blocking the sodium channel at 4%, calcium channels at 3.6% and potassium channel at 27.9%.

13.3.2.2 Binary classification

We start this section by describing the outcome of the greedy algorithm selection. These are shown, for the three classification problems addressed, in Figure 13.12, in which the weights of the dictionary entries are plotted. The selected entries are different (also in number) for the different classification problems. For instance, for the sodium binary classification, we obtained 3 components for the dimension 1 whereas for the same dimension, we obtained 4 components for the potassium case and 5 for the calcium case. In all the cases, the linear combinations retained are sparse.

The classification results are reported hereafter. First, an aggregated result is presented (considering all the different concentrations, providing an overall label). Then, in the last part of this section, the results at different concentrations are described.

Aggregated results Figure 13.13 shows classification results for the seven compounds that were included in the validation set. The value shown for each compound corresponds to the success rate of classifying the compound correctly as a blocker or non-blocker for either the sodium, potassium or calcium channel according to their label (see Table 13.1). The results for the 7 molecules in the validation set are commented.

1. ***Azimilide**: potassium, sodium and calcium channel blocker* (Table 13.1)

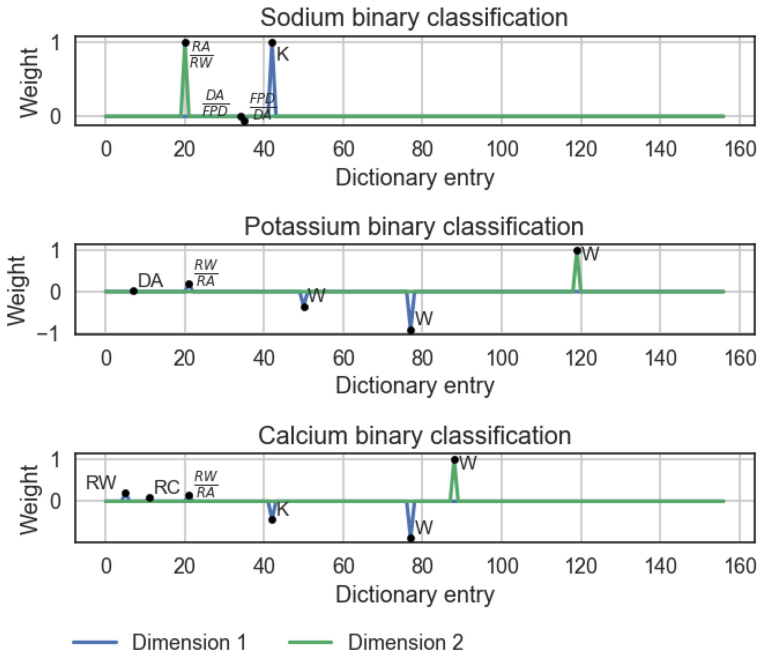

Figure 13.12: Binaray classification part): Weights obtained by the optimised classification algorithm.

Azimilide is well classified as a potassium channel blocker with a high confidence and for 90% of the sample. The sodium channel blockade by Azimilide is clearly not seen by the classifier as 90% of the sample is labelled as non-sodium channel blockade with a confidence close to 100%. The calcium channel blockade classification is also less clear as only 70% of the samples are labelled as non-calcium channel blockade with almost 80% of confidence. This could be related to the potency of Azimilide to block the inward sodium currents and L-type calcium channels is lower than for blocking the hERG channel [HQ06]. Besides, the highest concentration tested was lower than the $IC50$ values for blocking sodium and calcium channels (Table 13.1). A dictionary entry chosen by the algorithm for potassium and calcium blockade classification is the ratio $\frac{RW}{RA}$ (see Figure 13.12). As shown in [VZJ+18], hERG channel block can induce a T-wave flattening in the ECG. This phenomenon is also observed in the FP repolarisation of Pluricyte Cardiomyocytes for $0.1\mu M$ of Bepridil (see Figure 13.14) and could be an explanation of the $\frac{RW}{RA}$ selection by the algorithm.

2. ***Bepridil****: potassium, calcium and sodium channel blocker* (Table 13.1)
Sodium and potassium channel blockade classification for Bepridil is well captured by the classifier (with high proportion and high confidence). Calcium channel blockade is not seen by the classifier for Bepridil. A potential explanation could be that if calcium and potassium channels are blocked simultaneously, Bepridil does not show

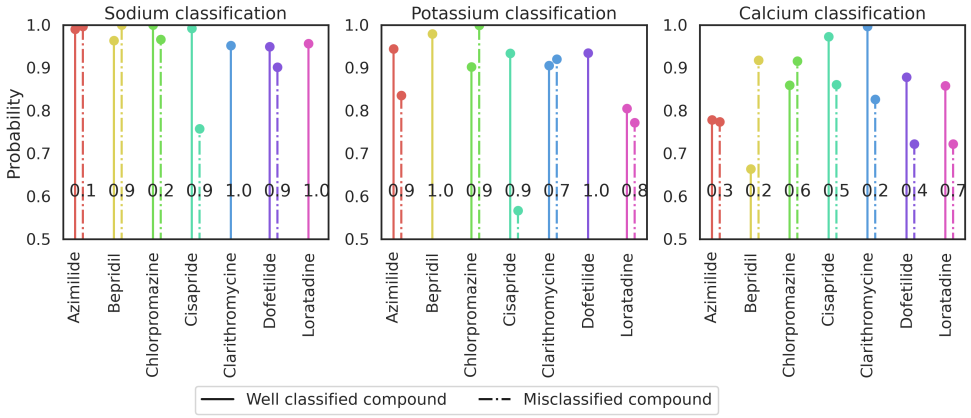

Figure 13.13: Binary classification part: Experimental data classification in binary case. Plain (resp. dotted) lines correspond to the average confidence (y-axis) of the LDA classifier for well classified (resp. misclassified) compound (well classification is according to Table 13.1). The black values on the lines correspond to the proportion of well classified observations for each compound.

a specific pattern of a calcium channel blocker, but essentially potassium and sodium channel patterns are detected as shown in Figure 13.14 (e.g. FPD prolongation due to potassium channel block and DA decrease due to sodium channel block).

3. ***Chlorpromazine**: potassium, calcium and sodium channel blocker* (Table 13.1)
   Chlorpromazine is well classified for the potassium and calcium channel classifications (i.e. it is considered as a potassium channel blocker and calcium channel blocker). The success rates for Chlorpromazine are similar to those obtained with Loratadine. An explanation might be the fact that they approximately have the same factor between the hERG and $Ca_v1.2$ $IC50$ values. In addition, Chlorpromazine is classified as a sodium channel blocker in only 20% of all cases, but with a probability of 100%.

4. ***Cisapride**: potassium channel blocker* (Table 13.1)
   If we compare classification results obtained for Chlorpromazine and Cisapride, the potassium channel binary classification success rate is the same. However, the classifier is more confident when Cisapride is well classified. Moreover, for the calcium channel classification, Chlorpromazine is classified as a calcium channel blocker in 60% of the cases whereas Cisapride is classified as non-calcium channel blocker in 50% of the cases with a higher confidence than when Cisapride is misclassified as calcium channel blocker. Moreover, in 90% of the samples tested, Cisapride is being classified as a non-sodium channel blocker with a confidence close to 100%.These results are in good agreement with the high potency of Cisapride to block the hERG channel, and the multi-channel block ability (hERG, $Na_v1.5$ and $Ca_v1.2$) for Chlorpromazine.

5. ***Clarithromycine**: potassium channel blocker* (Table 13.1)
   Clarithromycine is well considered as a non-sodium channel blocker with a high confidence and for all tested samples. In 70% of the cases Clarithromycine is well classified as a potassium channel blocker with around 90% of confidence. However, Clarithromycine is also labelled as a calcium blocker for 80% of the samples and with more than 80% of confidence. Important to note here is that, although Clarithromycine is labelled as a non-calcium channel blocker for only 20% of the samples, the confidence for this well classification is close to 100%.

6. ***Dofetilide**: potassium channel blocker* (Table 13.1)
   The classifier always returns Dofetilide as a potassium channel blocker with a high probability. Dofetilide is also classified as a sodium blocker, but only for 10% of the cases and with a lower probability than when it is not classified as a sodium blocker. For the calcium channel block classification, Dofetilide is considered as a non-calcium channel blocker for 40% of the cases but with a higher probability than when it is considered as a calcium channel blocker.

7. ***Loratadine**: potassium and calcium channel blocker* (Table 13.1)
   For the sodium channel block classification, Loratadine is always well classified (as a non-sodium channel blocker) with high confidence (averaged probability returned by the classifier is close to one, see Figure 13.13). For potassium blockade, Loratadine is well classified in 80% of the cases. Moreover, when Loratadine is well classified, the classifier is more confident (> 0.8) than when it is misclassified (< 0.8).

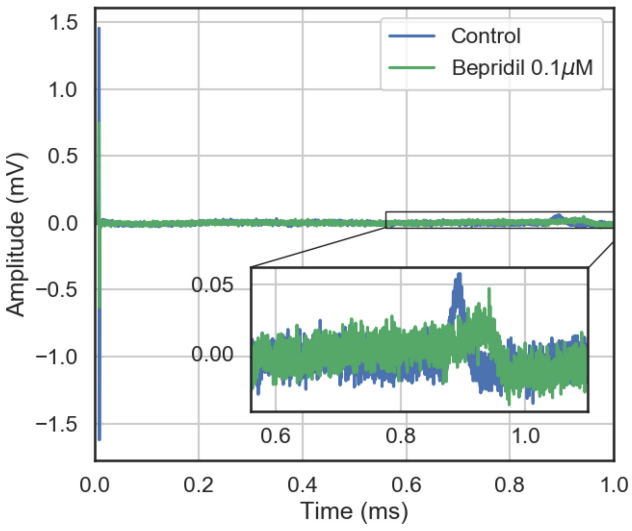

Figure 13.14: Binary classification part, Bepridil classification results: Example of experimental data with Bepridil, showing an increase in FPD and a decrease in DA of Pluricyte Cardiomyocytes.

For most of the cases, drugs are well classified with a high confidence. However, this is not always the case. For instance, Dofetilide has been perfectly classified as a potassium channel blocker with a high confidence (around 90%), but Dofetilide has also been misclassified as a calcium channel blocker with a high confidence (around 70%).

Study for each concentration Details for ion channel block classification of each concentration of each compound are given in Figure 13.15. This figure shows how each compound was classified at each concentration. The interest is to study the evolution of the classification with respect to increasing concentrations.

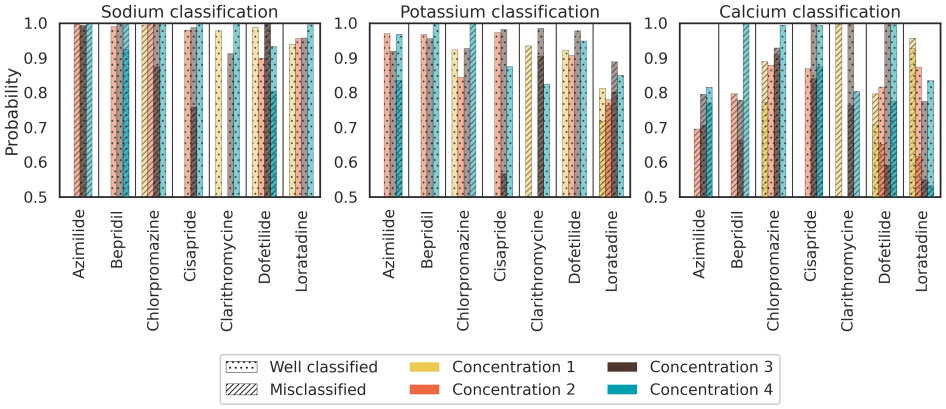

Figure 13.15: Binary classification part: Experimental data classification in binary case for each concentration. Some concentrations were not used due to the quiescence or noisy signal observation. For each concentration, the LDA classifier returns the average probability for well classified (dotted bars) and misclassified (hatched bars) compounds.

As done in the previous section, we present the results for each of the 7 molecules in the validation set.

1. ***Azimilide****: potassium, sodium and calcium channel blocker* (Table 13.1)
   Azimilide is classified as a potassium channel blocker with a probability higher than 90% for all concentrations tested. However Azimilide is misclassified for sodium and calcium channel blockade. As above-mentioned, this could be related to the fact that Azimilide blocks the inward sodium currents and L-type calcium channels at concentrations $5-10$ times higher than required for blocking the hERG channel [HQ06].

2. ***Bepridil****: potassium, calcium and sodium channel blocker* (Table 13.1)
   Bepridil is well classified as a sodium and potassium channel blocker with a high confidence. This is not the case for the calcium classification. An explanation could be that the potassium channel blockade hides the effect of the calcium channel blockade as above mentioned.

3. ***Chlorpromazine****: potassium, calcium and sodium channel blocker* (Table 13.1)
   Chlorpromazine is known to block sodium, potassium and calcium channels (see Table 13.1). Only for the first three concentrations, Chlorpromazine is clearly seen as a potassium channel blocker (Figure 13.15). The fourth and highest concentration show that sodium and calcium channels are affected instead of potassium. This is in line with the different potency of Chlorpromazine for the different ion channels: Chlorpromazine blocks hERG more potently than sodium or calcium (see Table 13.1).The calcium channel blockade is confirmed by the fact that well-classified confidence for calcium channel block increases with concentration, in addition to being well classified in 60% of all cases.

4. ***Cisapride****: potassium channel blocker* (Table 13.1)
   Well-classified confidence for Cisapride is always higher than misclassified confidence regardless of the concentration and, particularly for the sodium and potassium channel classifiers. This is in line with Cisapride being a very potent potassium blocker (see Table 13.1).

5. ***Clarithromycine****: potassium channel blocker* (Table 13.1)
   Clarithromycine is better classified as a potassium channel blocker at higher concentrations (higher confidence for the third concentration and no misclassification for the fourth concentration). Also the sodium classifier shows us that for all test concentrations, Clarithromycine is well classified as a non-sodium channel blocker. However, for any concentration, the calcium classifier does not give us satisfactory results, which means that Clarithromycine is wrongly classified as a calcium channel blocker.

6. ***Dofetilide****: potassium channel blocker* (Table 13.1)
   Dofetilide was wrongly labelled as a calcium channel blocker in 60% of the cases (see Figure 13.13). However, the well-classified confidence increases strongly with the concentration (see Figure 13.15), which means that the confidence of Dofetilide being a calcium channel blocker decreases when the concentration increases. The well-classified probability for the sodium channel (Dofetilide being a non-sodium channel blocker) and potassium channel (Dofetilide being a potassium channel blocker) is around 90% or even higher for all concentrations tested.

7. ***Loratadine****: potassium and calcium channel blocker* (Table 13.1)
   We know from Figure 13.13 that Loratadine is always classified as a non-sodium channel blocker. From Figure 13.15 we can conclude that the confidence of Loratadine being a non-sodium channel blocker increases with higher concentrations. In addition, Loratadine has also been classified as a potassium channel blocker in 80% of the cases (see Figure 13.13). Figure 13.15 shows that the classification is the best at the highest concentration (no misclassification), which can be explained by the relatively low test concentrations compared to the $IC50$ values (Table 13.1). Moreover, for the first two concentrations in the potassium channel classification, the confidence is higher when Loratadine is well classified than when Loratadine is

misclassified. A bad mark is the increase of the misclassified confidence for the first three concentrations. However, for the highest concentration tested, none of the samples were misclassified. Concerning the classification for the calcium channel, the success rate of Loratadine to be classified as a calcium channel blocker was 70% (Figure 13.13); and based on Figure 13.15 we can conclude that the misclassified confidence decreases strongly when the concentration increases. This is in line with the differences seen in $IC50$ values between hERG, $Ca_v1.2$ and $Na_v1.5$ (see Table 13.1).

13.3.2.3 Ternary classification

For the ternary classification we only considered one classifier but with three outputs: sodium, potassium and calcium channel blocker. Aggregated results for the ternary classification are presented in Fig 13.16.

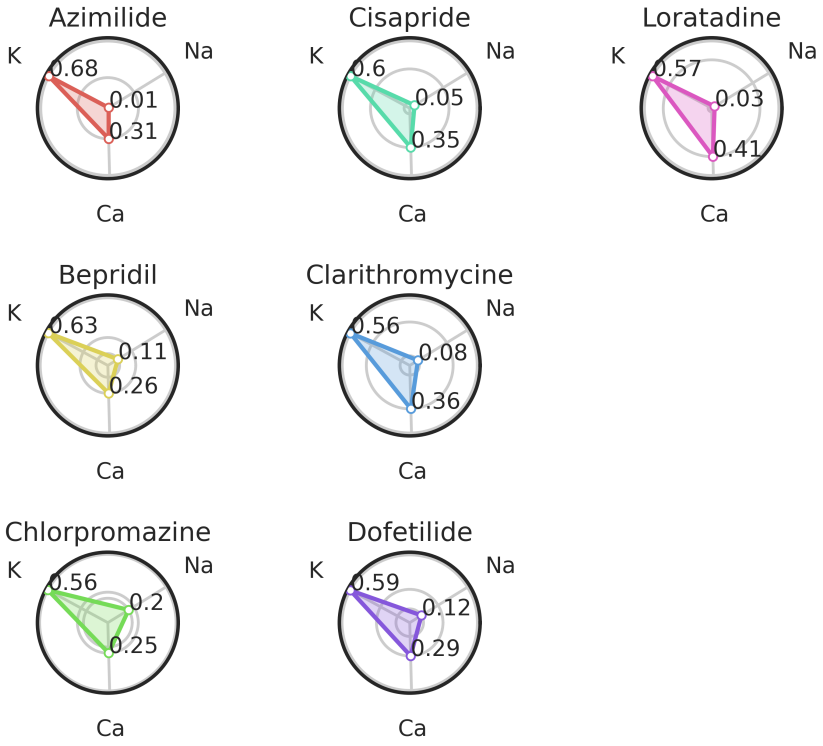

Figure 13.16: Ternary classification part: Experimental data classification in ternary case. Values returned by the classifier (black values in the polar plots) are the probabilities to block the corresponding channel blocker.

The ternary classifier classified all seven compounds from the validation set as potas-

sium channel blockers. As expected, the probability returned by the classifier decreases when the $IC50$ value increases (for example the probability for Loratadine to be a calcium channel blocker is 0.41 with $IC50 = 11.4\mu M$ (see Table 13.1) and the probability for Dofetilide to be a calcium channel blocker is 0.29 with $IC50 = 26.7\mu M$ (see Table 13.1)). These results do not take into account the different concentrations tested. The probabilities given for each concentration are given in Figure 13.17.

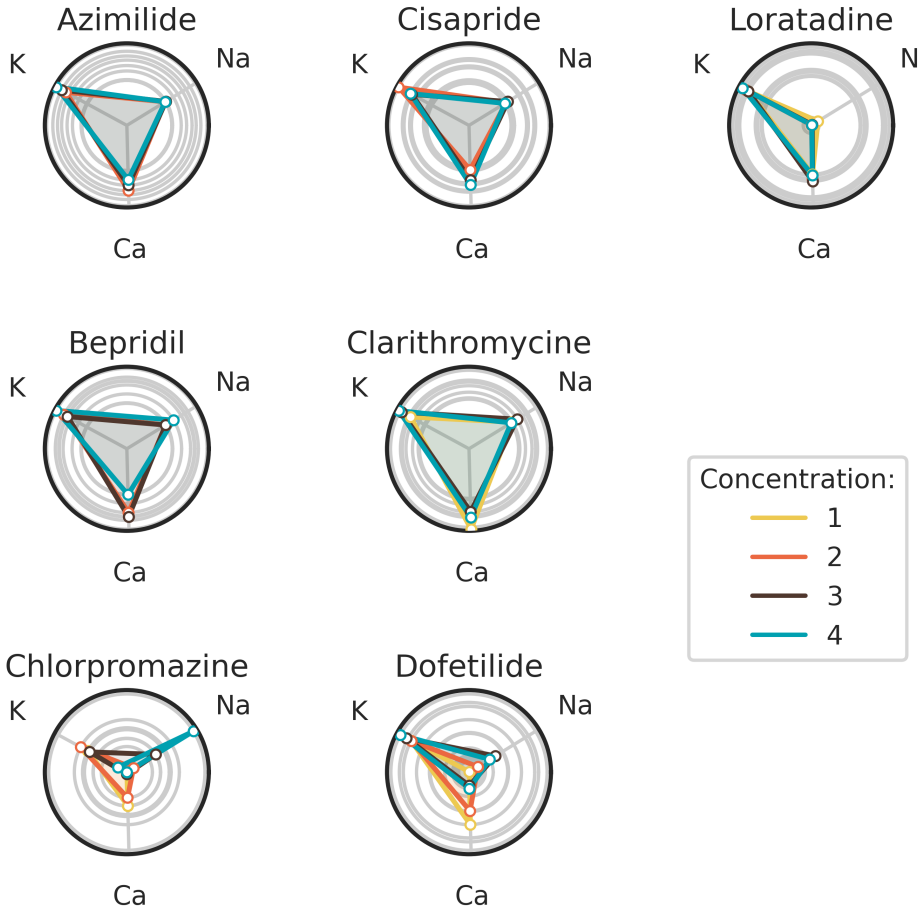

Figure 13.17: Ternary classification part: Experimental data classification in ternary case for each concentration.

The results for each of the 7 molecules are detailed hereafter. For sake of brevity, both aggregated and by-concentration results are commented.

1. ***Azimilide**: potassium, sodium and calcium channel blocker* (Table 13.1)
   Azimilide is well classified as a potassium blocker with a probability of 0.68 (see Figure 13.16). Although it is known that the potency of Azimilide to block the

inward sodium currents and L-type calcium channels is lower than blocking the hERG channel, the probability of being a sodium channel blocker was still lower than expected and did not change with higher concentrations.

2. ***Bepridil****: potassium, calcium and sodium channel blocker* (Table 13.1)
Bepridil is well classified as a potassium channel blocker with a probability equals to 0.63 (see Figure 13.16). The order of the different ion channel blockade probabilities is in good agreement with the $IC50$ values order (Table 13.1). The sodium channel blockade probability is 0.11. This probability is coherent in the sense that Bepidril is known to block sodium channel; other compounds which are not known as sodium channel blockers have a lower probability (0.01-0.08) of being a sodium channel blocker (except for Dofetilide at low concentrations). Unexpectedly, Figure 13.16 shows that the probability to be a calcium channel blocker is similar between Bepridil and Dofetilide (not a calcium channel blocker). Even for the last concentration of Bepridil, there is a decreasing confidence of being a calcium channel blocker in favour of being a potassium and sodium channel blocker (Fig 13.17). This could be explained by the fact that Bepridil has a higher potency for blocking hERG compared to blocking calcium channels and that the effects of hERG channel blockade masked the effects of blocking calcium channels.

3. ***Chlorpromazine****: potassium, calcium and sodium channel blocker* (Table 13.1)
Chlorpromazine is well classified as a potassium channel blocker with a probability equals to 0.56 (Figure 13.16). The probabilities for Chlorpromazine of being a calcium or sodium channel blocker are close to each other (Figure 13.16), which was expected as the $IC50$ values for calcium and sodium channel blockade are close to each other as well (Table 13.1). The confidence to classify Chlorpromazine as a sodium channel blocker is the highest for the highest concentration tested ($3\mu M$, see Table 13.1) (Figure 13.17). An explanation of this result could be that some compensation effects would appear on the repolarisation due to the simultaneous block of potassium as well as calcium. $3\mu M$ of Chlorpromazine corresponds at a 50% activity of the sodium channel (see conductance-block model in Section 13.2.2.2), which is clearly visible on the depolarisation amplitude (see Figure 13.18).

4. ***Cisapride****: potassium channel blocker* (Table 13.1)
Cisapride is well classified as a potassium channel blocker with a probability equals to 0.6 (see Figure 13.16). The second-highest confidence is for the calcium channel blocker. These channel blockade probabilities are in good agreement with the $IC50$ values of Cisapride for potassium channel blockade ($0.02\mu M$) and calcium channel blockade ($11.8\mu M$) (see Table 13.1). The difference in these values might also explain the observation that the confidence of being a potassium channel blocker decreases when the concentration increases, following by a higher confidence of Cisapride being a calcium channel blocker (Figure 13.17).

5. ***Clarithromycine****: potassium channel blocker* (Table 13.1)
Clarithromycine is well classified as a potassium channel blocker with a probability

equals to 0.56. It is interesting to see that the confidence of being a calcium channel blocker is lower for Clarithromycine than for Loratadine (see Figure 13.16). This point is expected because Loratadine is known to block calcium channels with an $IC50$ of $11.4\mu M$ (see Table 13.1), which is not the case for Clarithromycine ($IC50 > 30\mu M$). Another good point is that the confidence of being a potassium channel blocker for Clarithromycine slightly increases with higher concentrations (Figure 13.17).

6. ***Dofetilide**: potassium channel blocker* (Table 13.1)
   From Figure 13.16, we can see that Dofetilide is well classified as a potassium channel blocker with a probability equals to 0.59. If we now look at the classification results of Dofetilide in Figure 13.17, a high concentration gives a higher probability of being a potassium channel blocker and a lower probability to be a calcium channel blocker (which is in line with the binary classification method presented in Figure 13.14). This can be explained by the fact that for the three lower concentrations, the potassium activity is always higher than 90%(see 13.8 Table) whereas the highest concentration corresponds to a 75% activity (see 13.8 Table).

7. ***Loratadine**: potassium and calcium channel blocker* (Table 13.1)
   Loratadine is well classified as a potassium channel blocker with a probability equals to 0.57 (Figure 13.16). The second-highest probability concerns the calcium channel blocker, which was expected as Loratadine is more potent to block hERG ($6.1\mu M$, see Table 13.1) than to block the L-type calcium channel ($11.4\mu M$, see Table 13.1). Figure 13.17 shows us that the confidence of the classifier is almost the same regardless to the concentration.

13.4 Conclusion

Human iPSC-CMs are being increasingly adapted as a novel in vitro model to improve cardiac safety assessment. Of the many studies that have now investigated the impact of drugs on the electrophysiology of hiPSC-CMs, the most well-known is the multisite CiPA initiative. Data presented in [BDM+18] describes the utility of hiPSC-CMs in combination with MEA and voltage-sensing optical methods in evaluating the electrophysiological responses to 28 drugs linked to low, intermediate, and high TdP risk categories. Studies like the CiPA multisite study show promising results. However, predicting TdP risk at a reasonable level of accuracy remains a challenge. Besides, many screening platforms, like various MEA and calcium-flux devices, are becoming increasingly sophisticated and generate large multidimensional datasets. Improved automated analysis methods, including classification methods to accurately predict the risk for ion-channel block and TdP, are needed.

In the present work, a preliminary step towards the setup of high-throughput screening procedures was attempted. In particular, a method was proposed to systematically deal

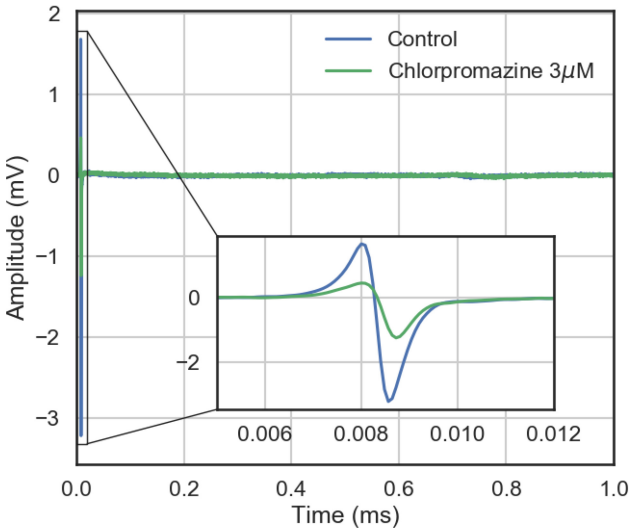

Figure 13.18: Ternary classification part, Chlorpromazine classification results: Example of experimental FP trace with Chlorpromazine.

with classification problems involving "CiPA" compounds for their risk to induce TdP as well as for their ion channel blocking properties.

13.4.1 Algorithm

This algorithm selects and combines pertinent features extracted from the signals in order to maximise the classification score (both in terms of the success rate and the confidence of the classifier) by means of a double greedy optimisation. The algorithm promotes sparsity (hence mitigating the overfitting risk) and it is fully scalable in terms of parallelism (remark that the number of cores can potentially equal the dictionary entry size). In this paper, the input space computed by the algorithm maximise a score by linearly separating the classes samples, using the classical LDA method. It would be interesting to test the algorithm with other classifiers such as support vector machine (SVM) with different kernels or k nearest neighbour (KNN) and against classification with PCA.

We applied the algorithm on simulated FP and calcium transient data for TdP risk classification as well as on *in vitro* data coming from FP signals recorded from hiPSC-CMs (Pluricyte Cardiomyocytes) that were cultured on 96 well MEA plates and subjected to 12 CiPA reference compounds (5 compounds were used as a training set and 7 were used to validate the algorithm).

13.4.2 TdP risk assessment

The classifiers obtained have given encouraging results for a drug safety profile of the compounds. Compounds known to have a high TdP risk were 100% well classified according to the arrhythmogenicity risk classification and a compound known to have a low TdP risk was well classified in 67% of the cases. This is conforming to the fact that we decided to put a strong weight on the type II error (wrongly classify a compound as non-torsadogenic). Concerning the torsadogenicity classification, more tests have to be done with higher concentrations (10xEFTPC, 50xEFTPC, ...). Thus, the compound impact on physiological traces (FP and intracellular calcium transient) would be more important, which would improve the classification (bigger margin between the data points and the separation plan). However, even at EFTPC, the TdP risk classification results are encouraging as only Propranolol was misclassified as torsadogenic. Particularly, the algorithm allows us to weight the type II error. To improve the arrhythmogenicity assessment, a ternary classifier could be established to distinguish low, moderate or high TdP risk.

13.4.3 Ion-channel blockade

Concerning ion-channel blockade classification of compounds, potassium was always well classified with a high confidence. Moreover, for the ternary classification study, for most of the tested compounds, the lowest the $IC50$ for a channel, the highest the confidence of the classifier to block this channel is.

The binary sodium channel blockade classification is good for all the compounds except for Chlorpromazine (at low concentration). The ternary classification study shows similar probabilities of Chlorpromazine for blocking the sodium channel as for blocking the calcium channel, which is in agreement with the similar $IC50$ values of Chlorpromazine for these channels.

However, the binary classification is less good for the calcium channel blockade classification. This could also be related to the fact that all CiPA compounds from the validation set block the hERG channel and with higher potency compared to blocking the calcium channel. The effect of blocking hERG could mask the effect of blocking calcium, making calcium channel blockade more difficult to classify.

In general, the binary and ternary classification strategies are in a good agreement (e.g. potassium channel blockade is always well classified). Nevertheless, more tests have to be done on the algorithm in order to validate and/or improve the classifiers.

For the channel block classification, simulations have been done only on highly pure channel block properties (no multi-channel blockade), simplified to only three types of channels: potassium, calcium or sodium, which is often not representative of the total ion channel blocking effects a compound could have. Training based on *in silico* multi-channel blockade would be more realistic and would most likely increase the robustness of the classification. Moreover, the present experimental protocol was performed at different concentrations for each compound. The dictionary entry could take this information into account.

Each application presented in this paper was based on one specific model of MEA device. It would also be interesting to know whether the MEA device might have an impact on the analysis of the drug effect, i.e. to study the case where we learn with one MEA device and we validate with data coming from another MEA device. The addition of intracellular calcium transient data would increase the classification in order to identify not only effects on ion-channels but also to detect negative and positive inotropic effects thereby having the capability to classify other classes of compounds, such as calcium-sensitizers or adrenergic receptor agonists.

The time compound dynamic was not studied in this paper. The dictionary could be extended with new biomarkers as beat rate or depolarisation standard deviation. These new entries could provide information on the impact of the compound on the monolayer stability. In order to represent this behaviour in the *in silico* dataset, a pacemaker action potential model showing experimental beat rate behaviour (Paci *et al*[PPC+18]) could be introduced.

In summary, the algorithm that we developed proved to be a promising tool to classify compounds for their risk to induce TdP as well as for their ion-channel blocking properties based on *in vitro* and in silico data derived from hiPSC-CMs. Therefore, this method can be implemented in *in vitro* MEA and/or calcium-flux studies using hiPSC-CMs where it may serve as a tool to improve machine learning approaches and to deliver fast and reliable prediction of drug-induced ion channel blockade and pro-arrhythmic behaviour to advance cardiac safety assessment.

13.5 Appendix

13.5.1 Field Potential Biomarkers computation

In this section, the computation of the biomarkers from FP time series is detailed. Let y be a FP signal. We defined as depolarisation part (t_1, y_1) from $t = 0$ to $t = 100ms$ and the repolarisation part (t_2, y_2) from $t = 100$ to $t = 1200ms$.
DA (Depolarisation Amplitude):

Difference between the maximum and minimum value of the FP during the depolarisation.

$$DA = max(y_1) - min(y_1).$$

RA (Repolarisation Amplitude):

Maximum in absolute value of the repolarisation.

$$RA = max(|y_2|).$$

FPD (Field Potential Duration):

Time difference between RA and the maximum in absolute value of the depolarisation.

For the depolarisation:

$$t_{dep} = t\left[\underset{t}{\text{argmax}}(|y_1(t)|)\right].$$

For the repolarisation:

$$t_{rep} = t\left[\underset{t}{\text{argmax}}(|y_2(t)|)\right].$$

Then,

$$FPD = t_{rep} - t_{dep}.$$

AUC_r (Area Under Curve of the repolarisation wave)

To get the repolarisation, y_2 is truncated around $\pm\Delta t$ of t_{rep}. We used $\Delta t = 100$ms. The trapezoidal rule is used to approximate the integral.

$$\text{AUCr} = \left|\int_{t_{rep}-\Delta t}^{t_{rep}+\Delta t} y_2(t)\mathrm{d}t\right|.$$

RC (Repolarisation Center)

Offset of the barycenter (with respect to time) of the repolarisation wave.

$$\text{RC} = \int_{t_{rep}-\Delta t}^{t_{rep}+\Delta t} t\bar{y}_2(t)\mathrm{d}t - t_{dep}.$$

With $\bar{y}_2(t)$ a rescaling such that it is strictly positive and integrates to 1 on $[t_{rep} - \Delta t, t_{rep} + \Delta t]$.

RW (Repolarisation Width)

Standard deviation of the repolarisation wave.

$$\mathrm{RW} = \left[\int_{t_{rep}-\Delta t}^{t_{rep}+\Delta t} t^2 \bar{y}_2(t) \mathrm{d}t - \left(\int_{t_{rep}-\Delta t}^{t_{rep}+\Delta t} t \bar{y}_2(t) \mathrm{d}t \right)^2 \right]^{1/2}.$$

FPN (FP Notch)

Potential value $4ms$ after t_{dep}. To be less sensitive to noise, the signal is multiplied by a test function $\phi(t_1) = \exp\left[-\frac{(t_1-(t_{dep}+4))^2}{0.04}\right]$.

$$\mathrm{FPN} = \int_{t_1} y_1(t_1)\phi(t_1)\mathrm{d}t_1.$$

13.5.2 Calcium Signals Biomarkers computation

In this section, the computation of the biomarkers from intracellular calcium concentration time series is detailed. Let y be the intracellular calcium concentration signal.

CA (Calcium Amplitude):

Difference between the maximum and minimum value of the signal.

$$CA = max(y) - min(y).$$

DC (Drowsing Calcium):

Corresponding to the resting calcium, computed as the minimum value of the signal.

$$DC = min(y).$$

CDX (Calcium Duration):

Similarly to APD, CDX is the time interval corresponding to $X\%$ repolarisation. Let denote by y_1 the signal from $t = 0ms$ to $t = t\left[\underset{t}{\mathrm{argmax}}(|y(t)|)\right] ms$ and y_2 the signal from $t = t\left[\underset{t}{\mathrm{argmax}}(|y(t)|)\right] ms$ to $t = 1200ms$.

For the depolarisation:

$$t_{dep} = t\left[\underset{t}{\mathrm{argmin}}(|y_1(t) - \frac{100-X}{100}CA + DC|)\right].$$

For the repolarisation:

$$t_{rep} = t\left[\underset{t}{\mathrm{argmin}}(|y_2(t) - \frac{100-X}{100}CA + DC|)\right].$$

Electrodes median	Electrodes mean	Electrodes max	Index	Entry name
0	7			DA
1	8			RA
2	9			FPD
3	10			AUC_r
4	11			RC
5	12			RW
6	13			FPN
14	22	30		RA/DA
15	23	31		DA/RA
16	24	32		RA/FPD
17	25	33		FPD/RA
18	26	34		DA/FPD
19	27	35		FPD/DA
20	28	36		RA/RW
21	29	37		RW/RA
			38	CD90
			39	CD75
			40	CD50
			41	CD25
			42	CA
			43	DC
			44	AUC90
			45	AUC75
			46	AUC50
			47	AUC25
48	49			CA*FPD
			50	CA*CD90
			51	CA*CD75
			52	CA*CD50
			53	CA*CD25
54	58			FPD*CD90
55	59			FPD*CD75
56	60			FPD*CD50
57	61			FPD*CD25
			62 to 66 / 38 to 42	K
			67 to 99 / 43 to 157	Wavelets

X: Specific to TdP risk study. X: Specific to channel study.

Table 13.5: Indices and names of the dictionary entries.

Then,

$$CDX = t_{rep} - t_{dep}.$$

Proof of Proposition 13.1.

In the perfect case, we have $y_*^{(i)} = \widehat{y}^{(i)}$ and $\widehat{p}^{(i)} = 1, \forall i \in \{1, \ldots, n_s\}$. Then,

$$\mu = -\frac{1}{n_s} \sum_{i=1}^{n_s} \Big[\frac{n_s}{n_1} \delta_1(\widehat{y}^{(i)}) \delta_1(y_*^{(i)}) + \frac{n_s}{n_{-1}} \delta_{-1}(\widehat{y}^{(i)}) \delta_{-1}(y_*^{(i)}) \Big].$$

$$\iff \mu = -\frac{1}{n_s} \sum_{i=1}^{n_s} \frac{n_s}{n_1} \delta_1(\widehat{y}^{(i)}) \delta_1(y_*^{(i)}) - \frac{1}{n_s} \sum_{i=1}^{n_s} \frac{n_s}{n_{-1}} \delta_{-1}(\widehat{y}^{(i)}) \delta_{-1}(y_*^{(i)}).$$

Moreover, we know that $y_*^{(i)} = 1$, n_1 times and $y_*^{(i)} = -1$, n_{-1} times. Finally, we find that the minimum value is $\mu = -2$. In the worst case, we have $y_*^{(i)} \neq \widehat{y}^{(i)}$ and $\widehat{p}^{(i)} = 1, \forall i \in \{1, \ldots, n_s\}$. Then,

$$\mu = -\frac{1}{n_s} \sum_{i=1}^{n_s} \Big[-\alpha \frac{n_s}{n_1} \delta_{-1}(\widehat{y}^{(i)}) \delta_1(y_*^{(i)}) - \frac{n_s}{n_{-1}} \delta_1(\widehat{y}^{(i)}) \delta_{-1}(y_*^{(i)}) \Big].$$

Moreover, we know that $y_*^{(i)} = 1$, n_1 times and $y_*^{(i)} = -1$, n_{-1} times. Then,

$$\mu = \frac{1}{n_s} \Big(\alpha \frac{n_s}{n_1} n_1 + \frac{n_s}{n_{-1}} n_{-1} \Big) \Longrightarrow \mu = 1 + \alpha.$$

■

Proof of Proposition 13.2.

We have the following cost function:

$$\mu = -\frac{1}{n_s} \sum_{i=1}^{n_s} \widehat{p}^{(i)} \left\{ \sum_{j=1}^{c} \Big[\frac{n_s}{n_j} \delta_j(\widehat{y}^{(i)}) \delta_j(y_*^{(i)}) - \frac{n_s}{n_j} \alpha_j \sum_{\substack{m=1 \\ m \neq j}}^{c} \delta_m(\widehat{y}^{(i)}) \delta_j(y_*^{(i)}) \Big] \right\}.$$

In the best case, we have $y_*^{(i)} = \widehat{y}^{(i)}$ and $\widehat{p}^{(i)} = 1, \forall i \in \{1, \ldots, n_s\}$. Then,

$$\mu = -\frac{1}{n_s} \sum_{i=1}^{n_s} \sum_{j=1}^{c} \frac{n_s}{n_j} \delta_j(\widehat{y}^{(i)}) \delta_j(y_*^{(i)}).$$

We know how many samples are labelled for each class (n_1 for class 1 and n_2 for class 2). Then,

$$\mu = -\frac{1}{n_s} \Big(\frac{n_s n_1}{n_1} + \frac{n_s n_2}{n_2} + \ldots + \frac{n_s n_c}{n_c} \Big) \Longrightarrow s_{n_s} = -c.$$

On the other hand, in the worst case, we have $y_*^{(i)} \neq \widehat{y}^{(i)}$ and $\widehat{p}^{(i)} = 1, \forall i \in \{1, \ldots, n_s\}$. Then,

$$\mu = -\frac{1}{n_s}\sum_{i=1}^{n_s}\left\{\sum_{j=1}^{c} -\frac{n_s}{n_j}\alpha_j \sum_{\substack{m=1\\ m\neq j}}^{c} \delta_m(\widehat{y}^{(i)})\delta_j(y_*^{(i)})\right\}.$$

$$\iff \mu = \frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1}^{c}\frac{n_s}{n_j}\alpha_j \sum_{\substack{m=1\\ m\neq j}}^{c} \delta_m(\widehat{y}^{(i)})\delta_j(y_*^{(i)}).$$

For the same reasons concerning the number of labels in each class, we have:

$$\mu = \frac{1}{n_s}\Big(\frac{n_s}{n_1}\alpha_1 n_1 + \frac{n_s}{n_2}\alpha_2 n_2 + \ldots + \frac{n_s}{n_c}\alpha_k n_c\Big) \Longrightarrow \mu = \sum_{j=1}^{c}\alpha_j.$$

■

drug	Na IC50 (nM)	Ca IC50 (nM)	K IC50 (nM)	EFTPC (nM)
Amiodarone	15900.0	1900.0	860.0	0.8
Astemizole*	3000.0	1100.0	4.0	0.3
Bepridil*	2300.0	1000.0	160.0	35.0
Ceftriaxone	555900.0	153800.0	445700.0	23170.0
Chlorpromazine*	3000.0	3400.0	1500.0	38.0
Cilostazol	93700.0	91200.0	13800.0	128.0
Cisapride*	337000.0	11800.0	20.0	3.0
Clozapine*	15100.0	3600.0	2300.0	71.0
Dasatinib	76300.0	81100.0	24500.0	41.0
Diazepam	306400.0	30500.0	53200.0	29.0
Diltiazem*	22400.0	760.0	13200.0	122.0
Disopyramide*	168400.0	1036700.0	14400.0	742.0
Dofetilide*	162100.0	26700.0	30.0	2.0
Donepezil	38500.0	34300.0	700.0	3.0
Droperidol*	22700.0	7600.0	60.0	16.0
Duloxetine	5100.0	2800.0	3800.0	16.0
Flecainide	6200.0	27100.0	1500.0	753.0
Halofantrine	331200.0	1900.0	380.0	172.0
Haloperidol	4300.0	1300.0	40.0	4.0
Ibutilide*	42500.0	62500.0	18.0	140.0
Lamivudine	1571400.0	54200.0	2054000.0	19540.0
Linezolid	2644500.0	105400.0	1147200.0	59110.0
Loratadine*	28900.0	11400.0	6100.0	0.4
Methadone	31800.0	37400.0	3500.0	507.0
Metronidazole	2073200.0	177900.0	1340200.0	187000.0
Mibefradil	5600.0	510.0	1700.0	12.0
Mitoxantrone	93500.0	22500.0	539400.0	225.0
Moxifloxacin	1112000.0	173000.0	86200.0	10960.0
Nifedipine*	88500.0	12.0	44000.0	8.0
Nilotinib	13300.0	17500.0	1000.0	172.0
Nitrendipine*	21600.0	25.0	24600.0	3.0
Paliperidone	109000.0	193900.0	780.0	69.0
Paroxetine	9800.0	3900.0	1900.0	14.0
Pentobarbital	2686000.0	299000.0	1433900.0	5171.0
Phenytoin	72400.0	21900.0	147000.0	4360.0
Pimozide*	1100.0	240.0	40.0	0.5
Piperacillin	2433800.0	1226000.0	3405100.0	1378000.0
Procainamide	746600.0	389500.0	272400.0	54180.0
Quinidine*	14600.0	6400.0	720.0	3237.0
Raltegravir	824200.0	246700.0	782800.0	7000.0
Ribavirin	2997500.0	622500.0	967000.0	27880.0
Risperidone*	43400.0	34200.0	260.0	2.0
Saquinavir	12100.0	1900.0	16900.0	130.0
Sertindole	6900.0	6300.0	33.0	2.0
Sitagliptin	1220800.0	147100.0	174700.0	442.0
Solifenacin	1500.0	4300.0	280.0	3.0
Sotalol*	7013900.0	193300.0	111400.0	14690.0
Sparfloxacin	2555000.0	88800.0	22100.0	1766.0
Sunitinib	16500.0	33400.0	1200.0	13.0
Telbivudine	1095200.0	713900.0	422700.0	19720.0
Terfenadine*	2000.0	930.0	50.0	9.0
Terodiline	7400.0	4800.0	650.0	145.0
Thioridazine	1400.0	3500.0	500.0	980.0
Verapamil*	32500.0	200.0	250.0	88.0
Voriconazole	1550500.0	414200.0	490900.0	7563.0

drug	Na IC50 (nM)	Ca IC50 (nM)	K IC50 (nM)	EFTPC (nM)
Ajmaline	8200.0	71000.0	1040.0	1500.0
Amiodarone	4800.0	270.0	30.0	0.5
Amitriptyline	20000.0	11600.0	3280.0	41.0
Bepridil*	3700.0	211.0	33.0	33.0
Chlorpromazine*	4300.0	nan	1470.0	38.0
Cibenzoline	7800.0	30000.0	22600.0	976.0
Cisapride*	14700.0	nan	6.5	4.9
Desipramine	1520.0	1709.0	1390.0	108.0
Diltiazem*	9000.0	450.0	17300.0	122.0
Diphenhydramine	41000.0	228000.0	5200.0	34.0
Dofetilide*	300000.0	60000.0	5.0	2.0
Fluvoxamine	39400.0	4900.0	3100.0	377.0
Haloperidol	7000.0	1700.0	27.0	3.6
Imipramine	3600.0	8300.0	3400.0	106.0
Mexiletine*	43000.0	100000.0	50000.0	4129.0
Mibefradil	980.0	156.0	1800.0	12.0
Nifedipine*	37000.0	60.0	275000.0	7.7
Nitrendipine*	36000.0	0.35	10000.0	3.02
Phenytoin	49000.0	103000.0	100000.0	4500.0
Pimozide*	54.0	162.0	20.0	1.0
Prenylamine	2520.0	1240.0	65.0	17.0
Propafenone	1190.0	1800.0	440.0	241.0
Propranolol	2100.0	18000.0	2828.0	26.0
Quetiapine	16900.0	10400.0	5800.0	33.0
Quinidine*	16600.0	15600.0	300.0	924.0
Risperidone*	102000.0	73000.0	150.0	1.81
Sertindole	2300.0	8900.0	14.0	1.59
Tedisamil	20000.0	nan	2500.0	85.0
Terfenadine*	971.0	375.0	8.9	9.0
Thioridazine	1830.0	1300.0	33.0	208.0
Verapamil*	41500.0	100.0	143.0	81.0

Table 13.7: Drugs known as torsadogenic (red) and non-torsadogenic (green) with their IC50 and EFTPC from Mirams *et al.* *: CiPA compound [CFG+16].

Compound	Concentration 1			Concentration 2			Concentration 3			Concentration 4		
	hERG	Cav1.2	Nav1.5	hERG	Cav1.2	Nav1.5	hERG	Cav1.2	Nav1.5	hERG	Cav1.2	Nav1.5
Azimilide	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Bepridil	94.1	99.0	99.6	61.5	90.9	95.8	13.8	50.0	69.7	1.6	9.1	18.7
Chlorpromazine	94.0	97.3	96.9	83.3	91.9	90.9	61.2	78.2	76.0	33.3	53.1	50.0
Cisapride	86.2	100	100	66.7	99.9	100	38.8	99.7	100	16.7	99.2	100
Clarithromycine	99.7	NA	NA	97.1	NA	NA	76.7	NA	NA	24.8	NA	NA
Clozapine	96.0	97.4	99.4	88.4	92.3	98.0	70.8	79.1	94.1	43.4	54.5	83.4
Diltiazem	99.9	98.7	100	99.2	88.4	99.6	93.0	43.2	95.7	56.9	7.1	69.1
Dofetilide	99.0	100	100	96.8	100	100	90.4	100	100	75.0	100	100
Droperidol	65.4	99.6	99.9	37.5	98.7	99.6	15.9	96.0	98.6	5.7	88.4	95.8
Ibutilide	99.4	100	100	94.7	100	100	64.3	100	100	15.3	99.8	99.8
Loratadine	100	100	100	100	100	100	99.8	99.9	100	99.5	99.7	99.9
Mexiletine	99.8	99.9	99.7	98.4	99.2	97.4	86.1	92.6	79.2	38.3	55.6	27.5

Table 13.8: Percentage of activity using a Hill coefficient equals to 1.

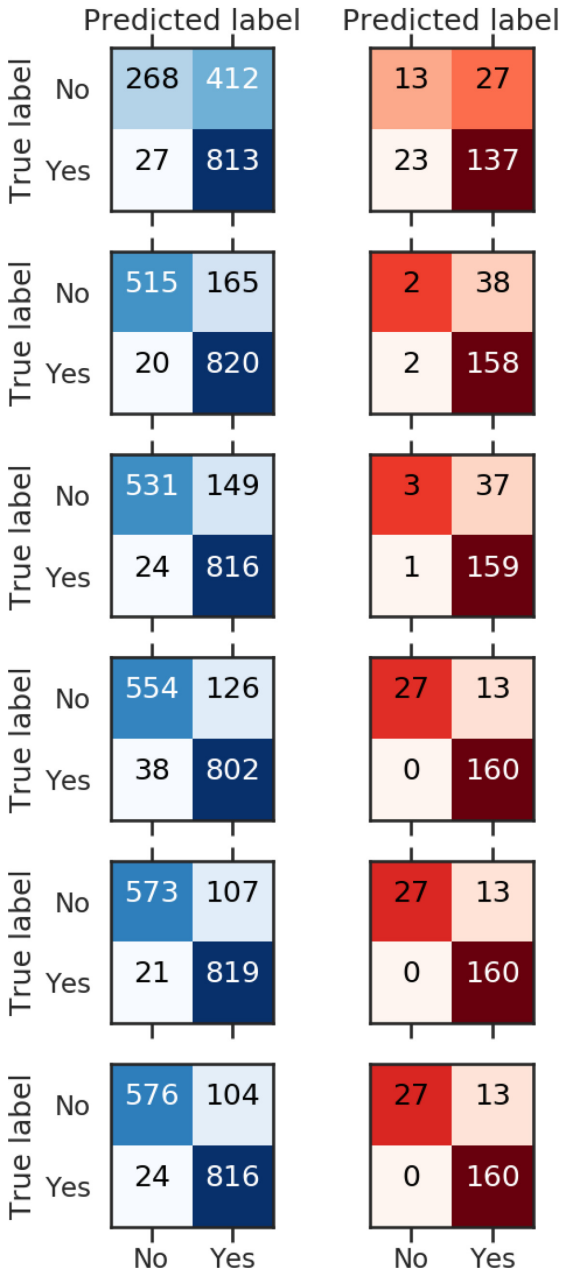

Figure 13.19: Confusion matrices obtained for TdP risk classification. From top to bottom we have respectively 1 component in $\mathbb{R}$, 2 components in $\mathbb{R}$, 3 components in $\mathbb{R}$, 3 components in $\mathbb{R}$ and 1 component in $\mathbb{R}^2$, 3 components in $\mathbb{R}$ and 2 components in $\mathbb{R}^2$ and 3 components in $\mathbb{R}$ and 3 components in $\mathbb{R}^2$. The left column corresponds to the training set and the right column to the validation set. No: No TdP risk. Yes: TdP risk.

Chapter 14

Application of ASE-HD/DGDR coupling on cardiac field potentials

While the first method (DGDR) consists in the construction of an oriented linear combination of the entries (see Section 5), the second method (ASE-HD) works on the samples by selecting the most pertinent in order to construct an augmented training set (see Section 6). In particular, the construction of the method was performed in such a way that a same score function is maximised and the same free parameter (number of nearest neighbours) has to be considered. This chapter aims at taking advantages of these two methodological works in order to improve classification performances.

In a first study we focused on the potassium channel blockade classification by considering the same experimental data used in the previous section (see Section 13). These data are cardiac field potential signals obtained through MEA devices and provided by Ncardia company (see Section 13.2.1 for more details on the experimental setup). The consideration of the two methods reaches to a success rate close to 0.98 instead of 0.89 in the previous section.

A second application on cardiac field potentials was performed in collaboration with Udo Kraushaar from NMI. A set of 28 compounds were available among which 14 were revealed to us.

Contents

14.1 Introduction

In the DGDR method (see Section 5), a training set (and validation set for the early stopping criterion) is needed to construct the goal oriented lower dimensional space. However, the learning process may fail for some reason (such as noise or wrongly labelled data). This main reasons led to the development of the ASE-HD method (see Section 6). Conversely, the ASE-HD method constructs an augmented training set but does not consider the dimension of the problem (it does not struggle with the curse of dimensionality). Table 14.1 summarises the main pros and cons of DGDR and ASE-HD methods.

Struggle with:	DGDR	ASE-HD
Curse of dimensionality	✓	✗
"Bad" data	✗	✓
Overfitting	✓*	✗

Table 14.1: Pros and cons of the DGDR and ASE-HD methods.*By consideration of the early stopping criterion on the validation set.

The main goal of this chapter is to couple these two methods by taking profit of them:

- Oriented dimension reduction (DGDR method).
- Selection of the most relevant samples (ASE-HD method).

The resulting augmented set is then in an as low dimensional space as possible (depending on the early stopping criterion) and maximises the classification success rate. In particular, the two methods were developed in such a way that the same score function has to be maximised. It means that, for both methods, by discretisation of the score function, the same free parameter k (number of nearest neighbours) has to be considered. See Section 14.2.1 for more details.

14.1.1 Applications

In this chapter two applications on MEA signals were performed:

1. The first application is devoted to the validation of the coupling strategy. To do this, we considered *in vitro* cardiac field potentials provided by Ncardia. In particular, we focus on the potassium channel blockade classification which led to a classification success rate close to 0.89.
2. The second application results in a collaboration with Udo Kraushaar from NMI who provided us the *in vitro* experimental data. This second application is also devoted to the potassium channel blockade classification. It aims at being closer to the real industrial application, through a larger set of compounds to study.

Details on the protocols, datasets and data processing are given hereafter.

14.2 Method

In this section we describe the strategy to combine the two methods and the classification strategy. Preprocessing specific to each study is described in the corresponding application.

14.2.1 Coupling

For this study, the coupling of the two methodological works is performed in the following order:

1. ASE-HD: construction of the augmented set.
2. DGDR: oriented dimension reduction.

The rationale is that we need a training set to perform the DGDR process. Then, once the ASE-HD method is performed the obtained augmented set is used as the training set of the DGDR method. The validation and test sets are the same for the two methods.

14.2.2 Post-processing

Sets being randomly generated, the construction followed by the ASE-HD/DGDR methods is repeated N times. The majority vote strategy is then performed (see Section 10.2.3). This process aims at improving the robustness of the classification and can be seen as a cross-validation step. Indeed, the ASE-HD method is optimised over a validation set, which may induce an overfitting (see Table 14.1 in the Introduction of this chapter). The cross-validation strategy coupled with the majority vote tends to overcome this overfitting risk.

14.3 First application: Ncardia dataset

14.3.1 Experimental setup

Some reminders are presented here. We referred to Section 13.2.1 for more details on the experimental setup. For this study experiments based on human induced pluripotent stem cells derived to cardiomyocytes (Pluricyte Cardiomyocytes) were performed on a 96 well Maestro MEA device provided by Axion BioSystems. From the 12 available compounds, a total of 7 compounds belong to the Test set whereas the 5 others belong to the reservoir or validation set. Experiments were repeated 5 times for each compound at a given concentration, the protocol consisting in one dose per well.

14.3.2 Pre-processing

The pre-processing part is divided into two phases. The first phase is devoted to the construction of the reservoir and validation sets to perform the ASE-HD method. The second phase consists in the rescaling of the data following the strategy defined below:

$$\begin{cases} \mu_i^{(0)} = \mathbb{E}[Tr_i^{(0)}], \mu_i^{(1)} = \mathbb{E}[Tr_i^{(1)}] \\ \sigma_i^{(0)} = \sigma(Tr_i^{(0)}), \sigma_i^{(1)} = \sigma(Tr_i^{(1)}) \\ M_i = \max(\mu_i^{(0)} + 2\sigma_i^{(0)}, \mu_i^{(1)} + 2\sigma_i^{(1)}), m_i = \min(\mu_i^{(0)} - 2\sigma_i^{(0)}, \mu_i^{(1)} - 2\sigma_i^{(1)}) \\ \begin{cases} Tr_i = \frac{Tr_i - m_i}{M_i - m_i} \\ Va_i = \frac{Va_i - m_i}{M_i - m_i} \\ Te_i = \frac{Te_i - m_i}{M_i - m_i} \end{cases} \end{cases},$$

where $S_i^{(j)}$ stands for the i^{th} dictionary of set $S \in \{Tr, Va, Te\}$ restricted to class $j \in \{0,1\}$, $\mathbb{E}[.]$ and $\sigma(.)$ denote the empirical mean and standard deviation of the considered set respectively. The choice of this strategy was already argued in Section 10.2.2.3.

14.3.2.1 Datasets construction

In addition of the *in vitro* experimental datasets, two kinds of sets were considered to either enrich the validation set either the reservoir:

- Augmented *in vitro* data: For the ASE-HD process, a statistical model was considered to enrich the validation set. For each entry, the empirical mean $\bar{\mu}$ and the empirical covariance matrix Σ is computed among the *in vitro* experimental data of each class (i.e potassium channel blocker or not). Then to generate the population, for each class we considered a gaussian distribution centred at $\bar{\mu}$ and with a covariance matrix equals to Σ. A total of 2500 data were generated with this method.

- *In silico* data: a total of 140 field potential simulations were used to build the reservoir. These *in silico* experiments are the same as the one used in the previous study (see Section 13.2.2 for more details on the simulations).

Then, the reservoir and validation sets were randomly generated: half of the *in vitro* experiments, *in silico* experiments and augmented *in vitro* data belongs to the reservoir whereas the other half belongs to the validation set. Then, each sample belongs to one and only one set. Finally, a rescale is performed on the whole samples following the process described in Section 10.2.2.3. Thus, most of the data are contained in the unit hypercube centred at $x = (\frac{1}{2}, \ldots, \frac{1}{2})$.

14.3.3 Results

Figure 14.1 shows success rates obtained with the majority vote strategy for different randomly constructed sets.

The combination of ASE-HD and DGDR reaches to the highest performances, with a success rate close to 0.98 (see ASE-HD+DGDR line in the figure). When we only consider the ASE-HD method (see ASE-HD line in the figure), the success rate is close to 0.87.

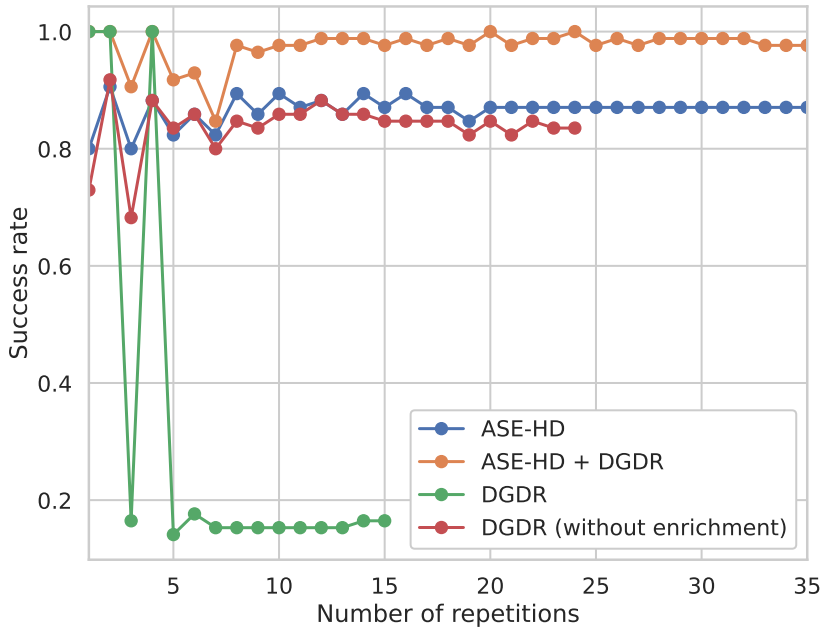

Figure 14.1: Cumulative success rate. The x-axis corresponds to the number of times the ASE-HD/DGDR methods are repeated. The y-axis corresponds to the success rate using the majority vote.

This score is quite similar to the 0.89 obtained in the original method (see Figure 13.13 in Section 13.3.2.2). Without the ASE-HD method, the DGDR method allows to obtain a score close to 0.16 (see DGDR line in the figure). This is essentially due to the enrichment process. Indeed, *in silico* experiments and data generated through the statistical model are imperfect and badly interfere the potassium channel blockade classification. This problem points out the drawback raised in Table 14.1. This is particularly obvious if compared with the case where the DGDR is performed without enrichment (see DGDR (without enrichment) in the figure). The ASE-HD method highlights that, among the enriched data, some of them are however relevant to improve performances of the classifier. The choice of the coupling strategy is therefore confirmed by the above results.

14.4 Second application: NMI dataset

14.4.1 Experimental setup

The MEA device used for the *in vitro* experiments was a plate with 96 wells and 3 electrodes per well provided by Multichannel Systems[1]. An example of one plate with

[1]Documentation available here

the corresponding protocol in each well is shown in Figure 14.2.

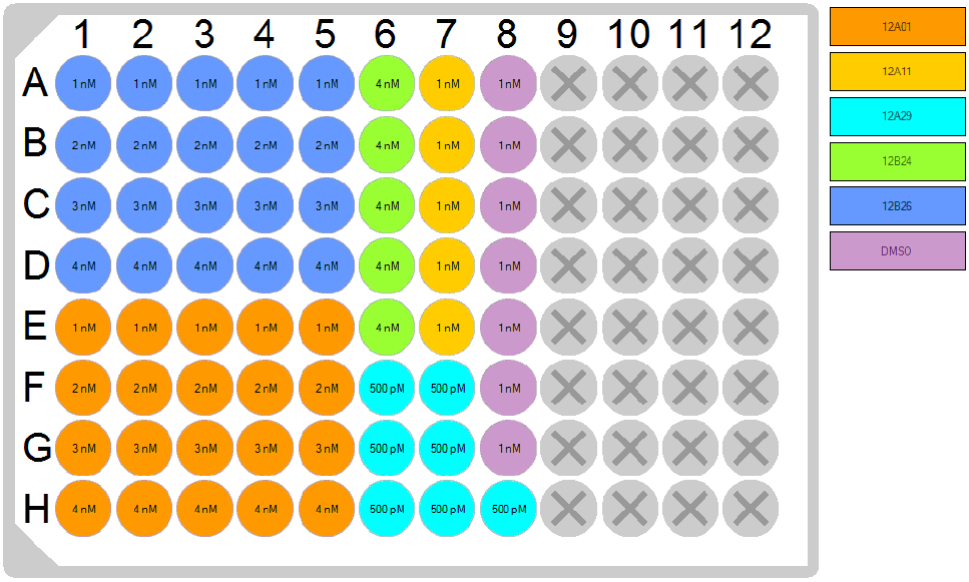

Figure 14.2: Scheme of one plate with the corresponding compound (code) and its concentration.

Cells used for the experiments are human induced pluripotent stem cells derived cardiomycytes provided by CDI[2] company. The MEA device used for the *in vitro* experiments is . The 28 compounds of the CiPA list were considered for the *in vitro* experiments. They are listed in Table 14.2.

A one dose per well scenario was performed in the experimental protocol. For each compound at a given concentration several replicates were realised (i.e same experimental scenario but in different wells of the MEA plate). The number of these replicates is also given in Table 14.2. Each well was coupled with a code corresponding to a specific compound. From the 28 CiPA compounds, 14 were revealed to us (see Table 14.2). These 14 compounds were then labelled and considered into the train, reservoir or the validation set. The 14 unrevealed compounds belong to the Test set.

An example of Field Potential signals recorded at one electrode at baseline and under compound addition ($3nM$ of Dofetilide) is shown in Figure 14.3.

14.4.2 Pre-processing

For this application, the pre-processing part is divided into three phases. The first phase consists in the construction fo the dictionary entry matrix. The second phase is devoted to the construction of the sets: Training, Validation and Test sets (and the

[2]Madison, USA. https://www.fujifilmcdi.com/.

Compound*	Concentration (nM) (Replicates)				Revealed	Antagonist to:
	#1	#2	#3	#4		
Astemizole	?	?	?	?	✗	K [HMEHhZ08]
Azimilide	?	?	?	?	✗	K [BEJ+98], Na [YT97], Ca [YT97]
Bepridil	1(5)	2(7)	3(5)	4(5)	✓	K [HMEHhZ08], Ca [YBS86], Na [YBS86]
Chlorpromazine	1(5)	2(5)	3(5)	4(5)	✓	K [HMEHhZ08], Ca [CVF84], Na [ONN89]
Cisapride	1(6)	2(5)	3(5)	4(5)	✓	K [HMEHhZ08]
Clarithromycine	?	?	?	?	✗	K [HMEHhZ08]
Clozapine	?	?	?	?	✗	K [LKK+06], Ca [NHW+17]
Diltiazem	1(5)	2(5)	3(5)	4(5)	✓	Ca [LT83]
Disopyramide	1(8)	2(7)	3(9)	4(10)	✓	Na [LXEB+20],K [HMEHhZ08]
Dofetilide	1(8)	2(5)	3(5)	4(5)	✓	K [HMEHhZ08]
Domperidone	?	?	?	?	✗	K [HMEHhZ08],Na [SVD+17]
Droperidol	1(5)	2(5)	3(8)	4(6)	✓	K [HMEHhZ08]
Ibutilide	1(5)	2(5)	3(5)	4(5)	✓	K [HMEHhZ08]
Loratadine	?	?	?	?	✗	K [Cru00], Ca [NHW+17]
Metoprolol	?	?	?	?	✗	β blocker [BDM+18]
Mexiletine	1(5)	2(5)	3(5)	4(5)	✓	Na [MWZ+13], K [GTBR+15]
Nifedipine	?	?	?	?	✗	Ca [ZWC+19]
Nitrendipine	1(5)	2(5)	3(5)	4(5)	✓	Ca [YB85], Na [YB85]
Ondansetron	?	?	?	?	✗	5-HT3 [TF92], K [CPP+17]
Pimozide	?	?	?	?	✗	K [HMEHhZ08],Ca [EDS+90]
Quinidine	1(5)	2(5)	3(5)	4(6)	✓	Na [Rod14], K [HMEHhZ08]
Ranolazine	1(5)	2(5)	3(5)	4(5)	✓	K [SZD+04], Na [BSF06]
Risperidone	1(5)	2(5)	3(5)	4(9)	✓	K [CWR05], Ca [CWR05]
D,I Sotalol	?	?	?	?	✗	K [HMEHhZ08]
Tamoxifen	?	?	?	?	✗	Na [HKKW03], K [HKKW03]
Terfenadine	?	?	?	?	✗	K [HMEHhZ08]
Vandetanib	?	?	?	?	✗	K [LHB+18], Na [LHB+18]
Verapamil	1(6)	2(7)	3(5)	4(7)	✓	Ca [ZWC+19]

* Colours corresponds to the TdP risk (green: low, orange: medium and red: high) [CFG+16].

Table 14.2: Experimental data information. The number of replicates corresponds to the number of wells on which the same experiment was performed.

reservoir when the ASE-HD method is performed). The last phase corresponds to the data rescaling and is performed in the same way as the strategy described in Section 14.3.2.

14.4.2.1 Dictionary entry computations

For this dataset, a sample is a couple of two beats, one corresponding to a control case and the second corresponding to a drug case (but for the same electrode of a same well in a same plate). A total of 18 quantities related to the depolarisation phase were extracted on the two beats. The first 6 quantities correspond to the depolarisation amplitude, duration, middle time of the depolarisation and its corresponding amplitude, the maximal slope and the minimal slope. The 6 others correspond to the average quantities over the whole beats of the trace (same plate, well and electrode). The last 6 quantities correspond to the standard deviation. The ratios with respect to the control, form the first 18 entries of the dictionary. An additional entry corresponds to the beat rate ratio between the drug case and the control case. This beat rate is computed on the full trace on which the

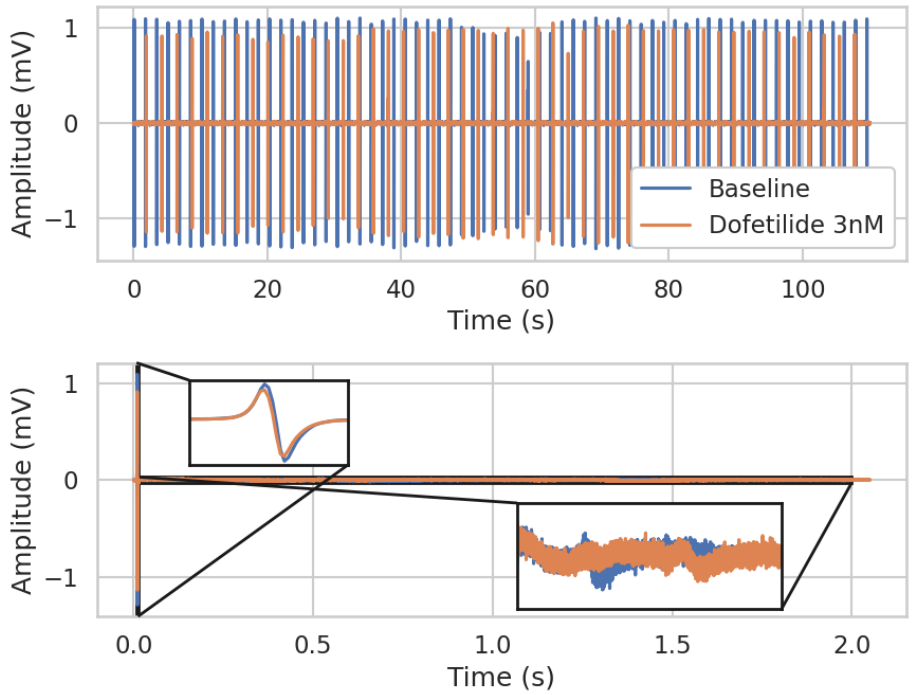

Figure 14.3: Example of recorded signals at one electrode. Upper: Traces at control case (baseline) and under compound addition for around 2 minutes of recordings. Traces are providing from the same electrode of the same well. Lower: Comparison between one random beat (from the full trace) at control case (baseline) and one random beat at $3nM$ of Dofetilide (from the full trace).

beat was extracted.

Then, 592 computed entries correspond to wavelet coefficients on the absolute difference between the two beats.

An example of signal and its reconstruction are shown in Figure14.5. The ℓ^2 error norm between the original signal and its reconstruction is approximately 0.49. This value corresponds to the limit value before being considered as an outlier in Figure 14.6, meaning that almost all the reconstructed signals have a better reconstruction that the one shown in Figure 14.5.

The last 592 entries are standard deviations of the wavelet coefficients computed on beats extracted on the same trace. It means that two samples providing from a same protocol (same compound, same concentration and same well) have the same last 592 entries. It results in a dictionary entry matrix of size $n_g = 1203$. The sample size n_s depends on the performed study cases shown in Section 14.4.3.

Remark 21

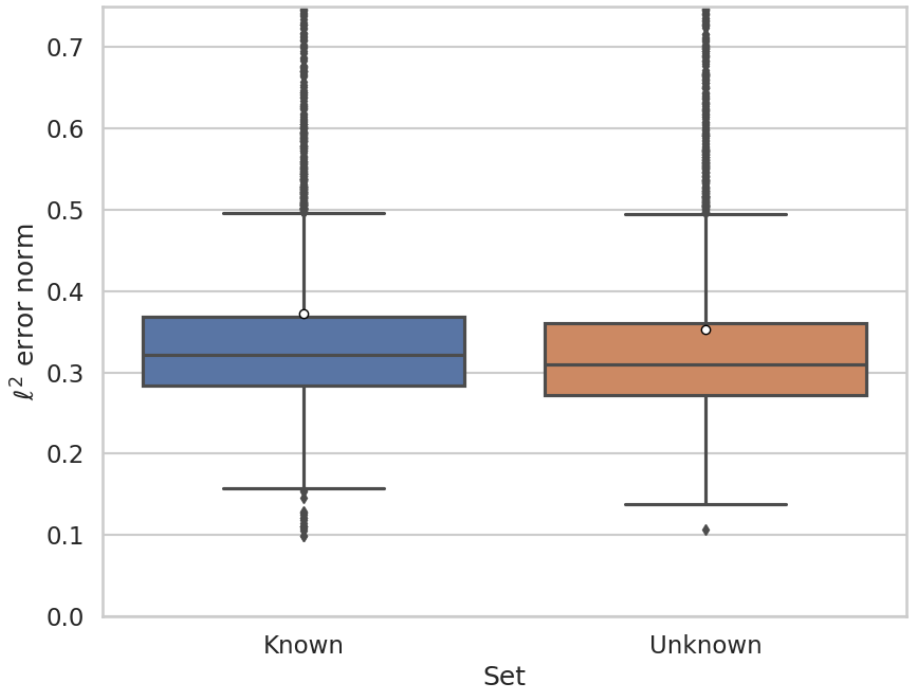

Figure 14.4: Boxplots of the ℓ^2 error norm between the signal and the reconstructed signal using wavelet coefficients. The 'Known' set corresponds to revealed compound whereas the 'Unknown' set corresponds to the unrevealed compounds (never used for the training and validation). See Table 14.2 for more details. White circles correspond to means.

These two beats come from the control case and under compound addition. However, in case of Hit/No Hit classification, the two beats can come from the control case.

14.4.2.2 Datasets construction and data rescaling

All the sets are randomly constructed in such a way that a sample belongs to only one set. Once the sets are constructed a data rescaling is performed with respect to the Training set using the same strategy as in the Section 14.3.2.

14.4.3 Results

14.4.3.1 Drug vs Control

The goal of this first study is to verify whether a compound at a given concentration has enough impact on the Field Potential to get spotted against control cases (without compound addition). This work is similar to the one previously realised in Section 10.3.1

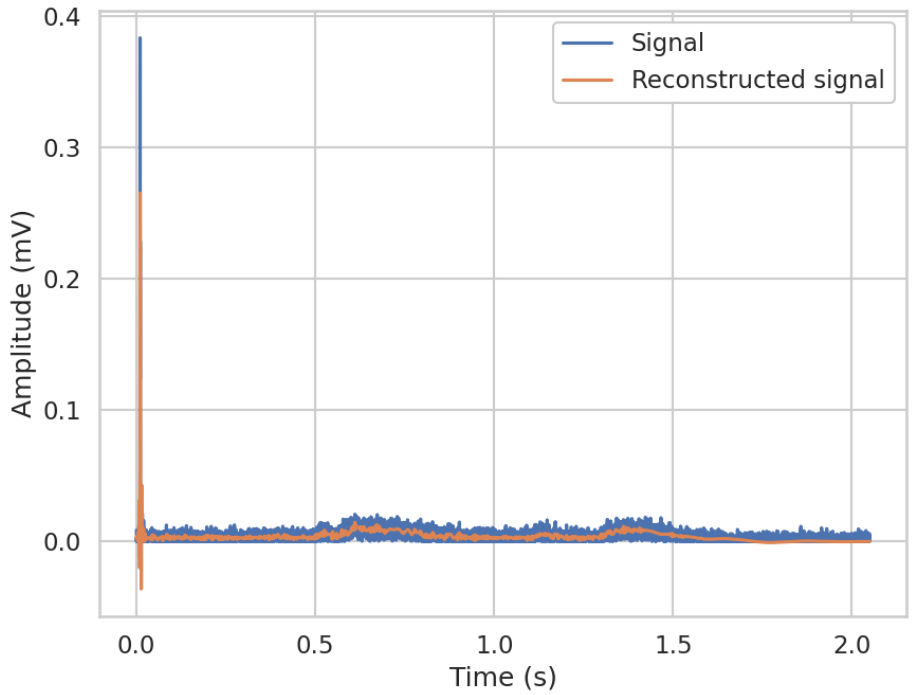

Figure 14.5: The signal to reconstruct corresponds to the absolute difference between the two beats shown in the above panel of Figure 14.3. The reconstruction led to an ℓ^2 error norm close to 0.49.

on automated patch clamp signals. This Hit/No Hit classification was performed on the 14 known compounds (see Table 14.2) at its lower concentration (except for the Dofetilide which is at $3nM$). To construct the Training, Validation and Test sets, a random process was repeated 10 times in such a way that each sample belongs to only one set. Moreover, the drug and control sample sizes are the same for each set. Finally, 60% of the total sample size is devoted to the Training set whereas 20% is devoted for the Validation set (and then 20% for the Test set). For this preliminary study only the DGDR method was considered. The classification success rates obtained on the Test sets are summarised in Figure 14.6.

All the compounds can easily be identified with an averaged success rate higher than 0.9. Moreover, the variability is globally low except for the Mexiletine (at $1nM$).

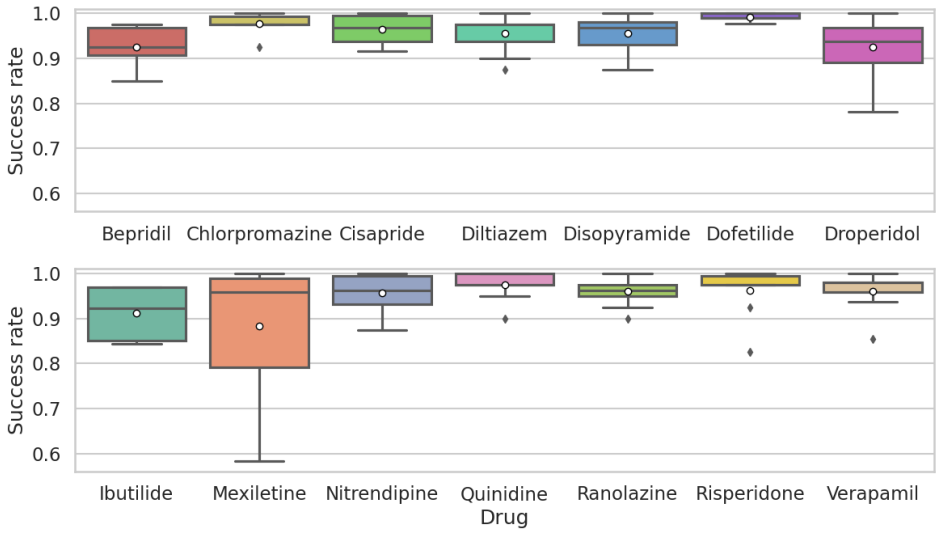

Figure 14.6: Hit/No Hit classification considering compounds at $1nM$ except for the Dofetilide at $3nM$. Boxplots are drawn over 10 processes to randomly construct the Training, Validation and Test sets. White circles correspond to means.

Chapter 15
Conclusions

MEA signals seems to carry enough information to perform an ionic channel blockade classification. In particular, we reached to a success rate close to 0.89 for the potassium channel blockade classification. The combined use of the DGDR and ASE-HD methods significantly improve the success rate to approximately 0.98.

The Hit/No Hit classification performed on the NMI dataset for the lowest concentrations led to a success rate higher than 0.9 for almost all the known compounds. Only Mexiletine has an averaged success rate lower than 0.9 while the median is around 0.96. It highlights that enough information is carried by the signal to discriminate control from a drug, meaning that the affected signal is higher than noise level at control case. Several investigations are in progress to study ion channel blockade classification.

Part V

Conclusion

This thesis aimed at proposing mathematical methods in order to provide a tool for classification problems in high dimensional-low sample size regimes, with applications in cardiac safety pharmacology. Each chapter of the manuscript ending with a dedicated conclusion, a global conclusion of the methodological part and its applications to cardiac safety pharmacology is presented hereafter.

Methodological aspects

This thesis proposed two methods addressing the high dimensional-low sample size regimes to solve classification tasks:

- DGDR: a double greedy dimension reduction method which allows to construct a sparse linear combination of entries in a "as low as" possible dimensional subspace. This dimension reduction is performed by maximising a score function related to the classification success rate and preventing the overfitting through an early stopping criterion.
- ASE-HD: an augmented set construction based on synthetic data generated by numerical simulations. The method relies on the Hausdorff distance between sets, considering the same score function as for the DGDR algorithm. It allows an automated rejection of samples considered as irrelevant to the considered classification problem.

The score function which has to be maximised allows the user to avoid the choice of a specific classifier and the ensuing hyperparameters (such as LDA or SVM). Its discretisation based on a K-Nearest Neighbours strategy depends on only one parameter which has to be tuned. Empirical studies performed in this manuscript suggest that the classification success rate obtained using the ASE-HD method is not very sensitive to the number of chosen nearest neighbours. However, a lower number of neighbours (e.g. 2 or 3) seems to reduce the computational time and increase the compression (meaning that fewer samples are needed to construct the augmented set). These aspects encourage the construction of a black box framework easy to use for the experimenter and satisfying the different constraints imposed by the industrial context:

- Deal with high dimension/low sample size regimes.
- Consider low/none assumptions on the observable space (i.e no a priori on the probability distributions).
- Exploit in silico models to improve the classification performances.
- Have few parameters/hyperparameters to be tuned.

Perspectives

Further works could either improve the proposed methods or open the door to other studies. These investigations are enumerated hereafter:

- Specific to the DGDR method:
  - Once the DGDR method is processed, computed weights are fixed. As in the online learning topic, in case of new experiments, some works could be done to tune the weights.

  - The dictionary entries could be the results of different classification methods (e.g. -1 or 1 in the case of a binary classification). The DGDR method will then consist of the construction of a lower subspace based on the contribution of the different classifiers. The DGDR method can therefore be seen in a sense as a boosting tool.

- Specific to the ASE-HD method:

  - By construction, the ASE-HD method tends to promote overfitting. A majority voting strategy was therefore considered to reduce it. However, a cross-validation strategy could be established, avoiding to repeat the DGDR process.
  - The extension of the ASE-HD method to regression problems was not studied in this manuscript. As for the DGDR method, an ℓ^2 norm could be considered instead of the score function in order to construct an augmented set for regression tasks.

- For the two methods:

  - Only binary classifications were considered in this thesis. An extension to more than two classes for the supervised classification problem could be implemented.
  - Extensions to semi-supervised and unsupervised classification problems can be investigated.
  - A probabilistic approach could be considered in the case where labelled data (or a part of them) are tagged with a confidence level.

Cardiac safety pharmacological aspects

In this thesis, cardiac safety pharmacology was the core of the applications raised by NOTOCORD®. These applications were made possible through several collaborations. They allowed to:

- Validate the methods: thanks to collaborations with Ncardia, NMI and Sophion companies.
- Extend the methods to regression problems: thanks to Esther Pueyo, David Adolfo Sampedro-Puente, Jesus Fernandez-Bes and Pablo Laguna, members of Zaragoza University.

Patch-clamp signal studies have shown particularly good results either for ionic channel activity estimation (on *in silico* experiments) or for the Hit/No Hit classification of the Nav1.7 channel (on *in vitro* experiments). As an example, for the first one, the DGDR method combined with UKF to estimate ion channel activities needs less than 5 cardiac cycles (beats) instead of decades for a same precision, which speeds up the process. For the second one, the DGDR method led to an accuracy close to 0.94 and increased the sensitivity from 0.47 (with the evaluation process performed by Sophion) to 0.52 (preserving a similar specificity).

The combined use of the DGDR and ASE-HD methods improved significantly the potassium ion channel blockade classification with a classification success rate close to 0.89 going up to 0.98.

Perspectives

Many works have to be done in different directions:

- Technological point of view:

  - The DGDR method was first implemented in Python. For slowness reasons appearing in high dimensions, a C/C++ version was implemented. The ASE-HD was only implemented in Python. For the same reasons, a C/C++ version should be more appropriate for an industrial application.
  - Implementation of a serviceable framework for pharmacologists. The actual version of the algorithms is not user-friendly and several manipulations have to be done to run the process (command line, paths specifications, ...).

- Drug development point of view:
  - Coupling between different physiological signals such as impedance or fluorescence may improve the compound analysis. This scenario is justified by the fact that some compounds could act on specific organites of the cells without affecting transmembrane channels (and then the electrical signal) but affecting the contractility (e.g. by blocking the sarcoplasmic reticulum).
  - *In silico* experiments could be improved. Existing models to simulate field potential signals consider a bi-dimensional resolution of the bidomain equations. This approximation could be not enough and some physical aspects could be considered (such as the layer between cardiac cells and the electrodes) to quantitatively improve *in silico* experiments.

Bibliography

[AB21] Tamim O Alabduladhem and Bruno Bordoni. Physiology, krebs cycle. *StatPearls [Internet]*, 2021.

[ABC+18] Emanuela Abbate, Muriel Boulakia, Yves Coudière, Jean-Frédéric Gerbeau, Philippe Zitoun, and Nejib Zemzemi. In silico assessment of the effects of various compounds in mea/hipsc-cm assays: Modeling and numerical simulations. *Journal of pharmacological and toxicological methods*, 89:59–72, 2018.

[AJD+15] Gudrun Antoons, Daniel M Johnson, Eef Dries, Demetrio J Santiago, Semir Ozdemir, Ilse Lenaerts, Jet DM Beekman, Marien JC Houtman, Karin R Sipido, and Marc A Vos. Calcium release near l-type calcium channels promotes beat-to-beat variability in ventricular myocytes from the chronic av block dog. *Journal of molecular and cellular cardiology*, 89:326–334, 2015.

[AJL+02] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. Principles of membrane transport. In *Molecular Biology of the Cell. 4th edition*. Garland Science, 2002.

[AKA91] David W Aha, Dennis Kibler, and Marc K Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[ASD21] Kalina Andrysiak, Jacek Stępniewski, and Józef Dulak. Human-induced pluripotent stem cell-derived cardiomyocytes, 3d cardiac structures, and heart-on-a-chip as tools for drug research. *Pflügers Archiv-European Journal of Physiology*, pages 1–25, 2021.

[ASM08] Babak Mohammadzadeh Asl, Seyed Kamaledin Setarehdan, and Maryam Mohebbi. Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal. *Artificial intelligence in medicine*, 44(1):51–64, 2008.

[AYY+17] Hiroyuki Ando, Takashi Yoshinaga, Wataru Yamamoto, Keiichi Asakura, Takaaki Uda, Tomohiko Taniguchi, Atsuko Ojima, Raku Shinkyo, Kiyomi Kikuchi, Tomoharu Osada, et al. A new paradigm for drug-induced torsadogenic risk assessment using human ips cell-derived cardiomyocytes. *Journal of pharmacological and toxicological methods*, 84:111–127, 2017.

[BBOV+17] Oliver J Britton, Alfonso Bueno-Orovio, Laszlo Virag, Andras Varro, and Blanca Rodriguez. The electrogenic na+/k+ pump is a key determinant of repolarization abnormality susceptibility in human ventricular cardiomyocytes: a population-based simulation study. *Frontiers in Physiology*, 8:278, 2017.

[BCDD14] Peter Binev, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Classification algorithms using adaptive partitioning. *The Annals of Statistics*, 42(6):2141–2163, 2014.

[BDM+18] Ksenia Blinova, Qianyu Dang, Daniel Millard, Godfrey Smith, Jennifer Pierson, Liang Guo, Mathew Brock, Hua Rong Lu, Udo Kraushaar, Haoyu Zeng, et al. International multisite study of human-induced pluripotent stem cell-derived cardiomyocytes for drug proarrhythmic potential assessment. *Cell reports*, 24(13):3582–3592, 2018.

[BDVR+03] SD Bird, PA Doevendans, MA Van Rooijen, A Brutel De La Riviere, RJ Hassink, R Passier, and CL Mummery. The human adult cardiomyocyte phenotype. *Cardiovascular research*, 58(2):423–434, 2003.

[BEJ+98] AE Busch, B Eigenberger, NK Jurkiewicz, JJ Salata, A Pica, H Suessbrich, and F Lang. Blockade of herg channels by the class iii antiarrhythmic azimilide: mode of action. *British journal of pharmacology*, 123(1):23–30, 1998.

[Bel15] Richard E Bellman. *Adaptive control processes*. Princeton university press, 2015.

[BF21] Damian C Bell and Bernard Fermini. Use of automated patch clamp in cardiac safety assessment: past, present and future perspectives. *Journal of Pharmacological and Toxicological Methods*, 110:107072, 2021.

[BGMW19] Louise A Bowler, David J Gavaghan, Gary R Mirams, and Jonathan P Whiteley. Representation of multiple cellular phenotypes within tissue-level simulations of cardiac electrophysiology. *Bulletin of mathematical biology*, 81(1):7–38, 2019.

[BGvdM92] Andrew R Barron, Lhszl Gyorfi, and Edward C van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE transactions on Information Theory*, 38(5):1437–1454, 1992.

[BH20] Caterina Bigoni and Jan S Hesthaven. Simulation-based anomaly detection and damage localization: an application to structural health monitoring. *Computer Methods in Applied Mechanics and Engineering*, 363:112896, 2020.

[BHS13] Aurélien Bellet, Amaury Habrard, and Marc Sebban. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.

[Big03] Brigitte Bigi. Using kullback-leibler distance for text categorization. In *European conference on information retrieval*, pages 305–319. Springer, 2003.

[Bis06] Christopher M Bishop. Pattern recognition. *Machine learning*, 128(9), 2006.

[BK20] Marcin Blachnik and Mirosław Kordos. Comparison of instance selection and construction methods with various classifiers. *Applied Sciences*, 10(11):3933, 2020.

[BKGW12] Paul W Burridge, Gordon Keller, Joseph D Gold, and Joseph C Wu. Production of de novo cardiomyocytes: human pluripotent stem cell differentiation and direct reprogramming. *Cell stem cell*, 10(1):16–28, 2012.

[BL13] Tom Burr and Claire Longo. Signal estimation using wavelet analysis of solution monitoring data for nuclear safeguards. *Axioms*, 2(1):44–57, 2013.

[Bla13] Axel Blau. Cell adhesion promotion strategies for signal transduction enhancement in microelectrode array in vitro electrophysiology: An introductory overview and critical discussion. *Current opinion in colloid & interface science*, 18(5):481–492, 2013.

[BM10] Anirban Basudhar and Samy Missoum. An improved adaptive sampling scheme for the construction of explicit boundaries. *Structural and Multidisciplinary Optimization*, 42(4):517–529, 2010.

[BMB08] Suhrid Balakrishnan, David B Madigan, and Peter Bartlett. Algorithms for sparse linear classifiers in the massive data setting. 2008.

[BOCF08] Alfonso Bueno-Orovio, Elizabeth M Cherry, and Flavio H Fenton. Minimal model for human ventricular action potentials in tissue. *Journal of theoretical biology*, 253(3):544–560, 2008.

[BPS+06] Dean Bottino, R Christian Penland, Andrew Stamps, Martin Traebert, Bérengère Dumotier, Anna Georgieva, Gabriel Helmlinger, and G Scott Lett. Preclinical cardiac safety assessment of pharmaceutical compounds using an integrated systems-based computer model of the heart. *Progress in biophysics and molecular biology*, 90(1-3):414–443, 2006.

[Bra05] Sarah K Branch. Guidelines from the international conference on harmonisation (ich). *Journal of pharmaceutical and biomedical analysis*, 38(5):798–805, 2005.

[BRGZ15] Muriel Boulakia, Fabien Raphel, Jean-Frédéric Gerbeau, and Philippe Zitoun. Toward transmembrane potential estimation from in vitro multi-electrode field potentials using mathematical modeling. *Journal of Pharmacological and Toxicological Methods*, (75):168–169, 2015.

[Bro07] Matthew Browne. A geometric approach to non-parametric density estimation. *Pattern Recognition*, 40(1):134–140, 2007.

[BSF06] L Belardinelli, JC Shryock, and H Fraser. Inhibition of the late sodium current as a potential cardioprotective principle: effects of the late sodium current inhibitor ranolazine. *Heart*, 92(suppl 4):iv6–iv14, 2006.

[BSV+17] Ksenia Blinova, Jayna Stohlman, Jose Vicente, Dulciana Chan, Lars Johannesen, Maria P Hortigon-Vinagre, Victor Zamora, Godfrey Smith, William J Crumb, Li Pang, et al. Comprehensive translational assessment of human-induced pluripotent stem cell derived cardiomyocytes for evaluating drug-induced arrhythmias. *Toxicological Sciences*, 155(1):234–247, 2017.

[BW93] G Bortolan and JL Willems. Diagnostic ecg classification based on neural networks. *Journal of Electrocardiology*, 26:75–79, 1993.

[CCG13] Dominique Chapelle, Annabelle Collin, and Jean-Frédéric Gerbeau. A surface-based electrophysiology model relying on asymptotic analysis and motivated by cardiac atria modeling. *Mathematical Models and Methods in Applied Sciences*, 23(14):2749–2776, 2013.

[CDD+12] Albert Cohen, Ingrid Daubechies, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Capturing ridge functions in high dimensions from point queries. *Constructive Approximation*, 35(2):225–243, 2012.

[CDM+17] Kelly C Chang, Sara Dutta, Gary R Mirams, Kylie A Beattie, Jiansong Sheng, Phu N Tran, Min Wu, Wendy W Wu, Thomas Colatsky, David G Strauss, et al. Uncertainty quantification reveals the importance of data variability and experimental design considerations for in silico proarrhythmia risk assessment. *Frontiers in physiology*, 8:917, 2017.

[CF20] Giuseppe C Calafiore and Giulia Fracastoro. Sparse ℓ^1-and ℓ^2-center classifiers. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[CFG+16] Thomas Colatsky, Bernard Fermini, Gary Gintant, Jennifer B Pierson, Philip Sager, Yuko Sekino, David G Strauss, and Norman Stockbridge. The comprehensive in vitro proarrhythmia assay (cipa) initiative—update on progress. *Journal of pharmacological and toxicological methods*, 81:15–20, 2016.

[CGB+16] Icilio Cavero, Jean-Michel Guillon, Veronique Ballet, Mike Clements, Jean-Frédéric Gerbeau, and Henry Holzgrefe. Comprehensive in vitro proarrhythmia assay (cipa): Pending issues for successful validation and implementation. *Journal of pharmacological and toxicological methods*, 81:21–36, 2016.

[CH14] Icilio Cavero and Henry Holzgrefe. Comprehensive in vitro proarrhythmia assay, a novel in vitro/in silico paradigm to detect ventricular proarrhythmic liability: a visionary 21st century initiative. *Expert opinion on drug safety*, 13(6):745–758, 2014.

[Che07] Karen C Cheung. Implantable microscale neural interfaces. *Biomedical microdevices*, 9(6):923–938, 2007.

[CHL05] José Ramón Cano, Francisco Herrera, and Manuel Lozano. Stratification for scaling up evolutionary prototype selection. *Pattern Recognition Letters*, 26(7):953–963, 2005.

[CJVJS16] William J Crumb Jr, Jose Vicente, Lars Johannesen, and David G Strauss. An evaluation of 30 clinical drugs against the comprehensive in vitro proarrhythmia assay (cipa) proposed ion channel panel. *Journal of pharmacological and toxicological methods*, 81:251–262, 2016.

[CN92] Michael Cahalan and Erwin Neher. [1] patch clamp techniques: An overview. *Methods in enzymology*, 207:3–14, 1992.

[CPP+17] Shanyu Cui, Hye-Won Park, Hye-Lim Park, Da-Som Mun, Hyo-Eun Kim, Nu-Ri Yun, and Bo-Young Joung. Ondansetron inhibits voltage-gated k+ current of ventricular myocytes from pregnant mouse. *International Journal of Arrhythmia*, 18(2):77–84, 2017.

[CPZ+09] Neil Castle, David Printzenhoff, Shannon Zellmer, Brett Antonio, Alan Wickenden, and Christopher Silvia. Sodium channel inhibitor drug discovery using automated high throughput electrophysiology platforms. *Combinatorial chemistry & high throughput screening*, 12(1):107–122, 2009.

[CRN98] Marc Courtemanche, Rafael J Ramirez, and Stanley Nattel. Ionic mechanisms underlying human atrial action potential properties: insights from a mathematical model. *American Journal of Physiology-Heart and Circulatory Physiology*, 275(1):H301–H321, 1998.

[CRT06] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.

[Cru00] William J Crumb. Loratadine blockade of k+ channels in human heart: comparison with terfenadine under physiological conditions. *Journal of Pharmacology and Experimental Therapeutics*, 292(1):261–264, 2000.

[CSC15] Eugene TY Chang, Mark Strong, and Richard H Clayton. Bayesian sensitivity analysis of a cardiac cell model using a gaussian process emulator. *PloS one*, 10(6):e0130252, 2015.

[Csi64] Imre Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten. *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, 8:85–108, 1964.

[CVF84] Brian K Chamberlain, P Volpe, and Sidney Fleischer. Inhibition of calcium-induced calcium release from purified cardiac sarcoplasmic reticulum vesicles. *Journal of Biological Chemistry*, 259(12):7547–7553, 1984.

[CWA+16] Chris Chambers, Ian Witton, Cathryn Adams, Luke Marrington, and Juha Kammonen. High-throughput screening of nav1. 7 modulators using a giga-seal automated patch clamp instrument. *Assay and drug development technologies*, 14(2):93–108, 2016.

[CWL+15] Po-Cheng Chang, Hung-Ta Wo, Hui-Ling Lee, Shien-Fong Lin, Ming-Shien Wen, Yen Chu, San-Jou Yeh, and Chung-Chuan Chou. Role of sarcoplasmic reticulum calcium in development of secondary calcium rise and early afterdepolarizations in long qt syndrome rabbit model. *PloS one*, 10(4):e0123868, 2015.

[CWR05] Torsten Christ, Erich Wettwer, and Ursula Ravens. Risperidone-induced action potential prolongation is attenuated by increased repolarization reserve due to concomitant block of i ca, l. *Naunyn-Schmiedeberg's archives of pharmacology*, 371(5):393–400, 2005.

[DBP+08] John Dunlop, Mark Bowlby, Ravikumar Peri, Dmytro Vasilyev, and Robert Arias. High-throughput electrophysiology: an emerging paradigm for ion-channel screening and physiology. *Nature reviews Drug discovery*, 7(4):358–368, 2008.

[DCB+17] Sara Dutta, Kelly C Chang, Kylie A Beattie, Jiansong Sheng, Phu N Tran, Wendy W Wu, Min Wu, David G Strauss, Thomas Colatsky, and Zhihua Li. Optimization of an in silico cardiac cell model for proarrhythmia risk assessment. *Frontiers in Physiology*, 8:616, 2017.

[DCZ+13] Lutz Dümbgen, Del Conte-Zerial, et al. On low-dimensional projections of high-dimensional distributions. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pages 91–104. Institute of Mathematical Statistics, 2013.

[Den06] Hans-Werner Denker. Potentiality of embryonic stem cells: an ethical problem even with alternative stem cell sources. *Journal of Medical Ethics*, 32(11):665–671, 2006.

[DET05] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.

[DHM20] Antoine Dedieu, Hussein Hazimeh, and Rahul Mazumder. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. *arXiv preprint arXiv:2001.06471*, 2020.

[DL97] Manoranjan Dash and Huan Liu. Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156, 1997.

[DL00] Manoranjan Dash and Huan Liu. Feature selection for clustering. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 110–121. Springer, 2000.

[DMJRM00] Roy De Maesschalck, Delphine Jouan-Rimbaud, and Désiré L Massart. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, 50(1):1–18, 2000.

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square hellinger subadditivity for bayesian networks and its applications to identity testing. In *Conference on Learning Theory*, pages 697–703. PMLR, 2017.

[ECKT17] David A Eisner, Jessica L Caldwell, Kornél Kistamás, and Andrew W Trafford. Calcium and excitation-contraction coupling in the heart. *Circulation research*, 121(2):181–195, 2017.

[EDS+90] JJ Enyeart, RT Dirksen, VK Sharma, DJ Williford, and SS Sheu. Antipsychotic pimozide is a potent ca2+ channel blocker in heart. *Molecular pharmacology*, 37(5):752–757, 1990.

[FBB02] Niels Fertig, Robert H Blick, and Jan C Behrends. Whole cell patch clamp recording performed on a planar glass chip. *Biophysical journal*, 82(6):3056–3062, 2002.

[FHAG+16] Bernard Fermini, Jules C Hancox, Najah Abi-Gerges, Matthew Bridgland-Taylor, Khuram W Chaudhary, Thomas Colatsky, Krystle Correll, William Crumb, Bruce Damiano, Gul Erdemli, et al. A new perspective in the field of cardiac safety testing through the comprehensive in vitro proarrhythmia assay paradigm. *Journal of biomolecular screening*, 21(1):1–11, 2016.

[FHT03] Alexei A Fedotov, Peter Harremoës, and Flemming Topsoe. Refinements of pinsker's inequality. *IEEE Transactions on Information Theory*, 49(6):1491–1498, 2003.

[Fit61] Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445–466, 1961.

[FKF+92] Tetsushi Furukawa, Shinichi Kimura, Nanako Furukawa, AL Bassett, and RJ Myerburg. Potassium rectifier currents differ in myocytes of endocardial and epicardial origin. *Circulation research*, 70(1):91–103, 1992.

[Fod02] Imola K Fodor. A survey of dimension reduction techniques. Technical report, Lawrence Livermore National Lab., CA (US), 2002.

[Fry77] MJ Fryer. A review of some non-parametric methods of density estimation. *IMA Journal of Applied Mathematics*, 20(3):335–354, 1977.

[FSV12] Massimo Fornasier, Karin Schnass, and Jan Vybiral. Learning functions of few arbitrary linear parameters in high dimensions. *Foundations of Computational Mathematics*, 12(2):229–262, 2012.

[FYK+20] Hiroyuki Fukushima, Miki Yoshioka, Masahide Kawatou, Víctor López-Dávila, Masafumi Takeda, Yasunari Kanda, Yuko Sekino, Yoshinori Yoshida, and Jun K Yamashita. Specific induction and long-term maintenance of high purity ventricular cardiomyocytes from human induced pluripotent stem cells. *PloS one*, 15(11):e0241287, 2020.

[GA18] Mengyang Gu and Kyle Anderson. Calibration of imperfect mathematical models by multiple sources of data with measurement bias. *arXiv preprint arXiv:1810.11664*, 2018.

[GE03] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[Ges89] David B Geselowitz. On the theory of the electrocardiogram. *Proceedings of the IEEE*, 77(6):857–876, 1989.

[GF15] Necla Gunduz and Ernest Fokoué. Robust classification of high dimension low sample size data. *arXiv preprint arXiv:1501.00592*, 2015.

[GH08] Steven Givant and Paul Halmos. *Introduction to Boolean algebras*. Springer Science & Business Media, 2008.

[GLG+15] Kristin H Gilchrist, Gregory F Lewis, Elaine A Gay, Katelyn L Sellgren, and Sonia Grego. High-throughput cardiac safety evaluation and multi-parameter arrhythmia profiling of cardiomyocytes using microelectrode arrays. *Toxicology and applied pharmacology*, 288(2):249–257, 2015.

[GLL+11] Donglin Guo, Que Liu, Tengxian Liu, Gary Elliott, Mireille Gingras, Peter R Kowey, and Gan-Xin Yan. Electrophysiological properties of hbi-3000: a new antiarrhythmic agent with multiple-channel blocking properties in human ventricular myocytes. *Journal of cardiovascular pharmacology*, 57(1):79–85, 2011.

[GM83] David B Geselowitz and WT Miller. A bidomain model for anisotropic cardiac muscle. *Annals of biomedical engineering*, 11(3-4):191–206, 1983.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[GPB10] Eleonora Grandi, Francesco S Pasqualini, and Donald M Bers. A novel computational model of the human ventricular action potential and ca transient. *Journal of molecular and cellular cardiology*, 48(1):112–121, 2010.

[GS02] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[GSS19] Hyukjun Gweon, Matthias Schonlau, and Stefan H Steiner. The k conditional nearest neighbor algorithm for classification and class probability estimation. *PeerJ Computer Science*, 5:e194, 2019.

[GTBR+15] Roberta Gualdani, Francesco Tadini-Buoninsegni, Mariagrazia Roselli, Ivana Defrenza, Marialessandra Contino, Nicola Antonio Colabufo, and Giovanni Lentini. Inhibition of herg potassium channel by the antiarrhythmic agent mexiletine and its metabolite m-hydroxymexiletine. *Pharmacology research & perspectives*, 3(5):e00160, 2015.

[GVdWvdL+07] David J Gallacher, André Van de Water, Henk van der Linde, An N Hermans, Hua Rong Lu, Rob Towart, and Paul GA Volders. In vivo mechanisms precipitating torsades de pointes in a canine model of drug-induced long-qt1 syndrome. *Cardiovascular research*, 76(2):247–256, 2007.

[HAC+13] Kate Harris, Mike Aylott, Yi Cui, James B Louttit, Nicholas C McMahon, and Arun Sridhar. Comparison of electrophysiological data from human-induced pluripotent stem cell–derived cardiomyocytes to functional preclinical safety assays. *toxicological sciences*, 134(2):412–426, 2013.

[Har68] Peter Hart. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 14(3):515–516, 1968.

[Haw04] Douglas M Hawkins. The problem of overfitting. *Journal of chemical information and computer sciences*, 44(1):1–12, 2004.

[HBG+05] Alan C Hindmarsh, Peter N Brown, Keith E Grant, Steven L Lee, Radu Serban, Dan E Shumaker, and Carol S Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS)*, 31(3):363–396, 2005.

[HBI+18] Bence Hegyi, Tamás Bányász, Leighton T Izu, Luiz Belardinelli, Donald M Bers, and Ye Chen-Izu. β-adrenergic regulation of late na+ current during cardiac action potential is mediated by both pka and camkii. *Journal of molecular and cellular cardiology*, 123:168–179, 2018.

[HBT+10] Martin Hinterseer, Britt-Maria Beckmann, Morten B Thomsen, Arne Pfeufer, Michael Ulbrich, Moritz F Sinner, Siegfried Perz, H-Erich Wichmann, Csaba Lengyel, Rainer Schimpf, et al. Usefulness of short-term variability of qt intervals as a predictor for electrical remodeling and proarrhythmia in patients with nonischemic heart failure. *The American journal of cardiology*, 106(2):216–220, 2010.

[HH52] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500–544, 1952.

[HJP03] Peg Howland, Moongu Jeon, and Haesun Park. Structure preserving dimension reduction for clustered text data based on the generalized singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):165–179, 2003.

[HKC+04] Dave Higdon, Marc Kennedy, James C Cavendish, John A Cafeo, and Robert D Ryne. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, 26(2):448–466, 2004.

[HKKW03] Jianying He, Margaret E Kargacin, Gary J Kargacin, and Christopher A Ward. Tamoxifen inhibits na+ and k+ currents in rat ventricular myocytes. *American Journal of Physiology-Heart and Circulatory Physiology*, 285(2):H661–H668, 2003.

[HLLS20] Minki Hwang, Chul-Hyun Lim, Chae Hun Leem, and Eun Bo Shim. In silico models for evaluating proarrhythmic risk of drugs. *APL bioengineering*, 4(2):021502, 2020.

[HMEHhZ08] Jules C Hancox, Mark J McPate, Aziza El Harchi, and Yi hong Zhang. The herg potassium channel and herg screening for drug-induced torsades de pointes. *Pharmacology & therapeutics*, 119(2):118–132, 2008.

[HML+03] Jia-Qiang He, Yue Ma, Youngsook Lee, James A Thomson, and Timothy J Kamp. Human embryonic stem cells develop into multiple types of cardiac myocytes: action potential characterization. *Circulation research*, 93(1):32–39, 2003.

[HMN05] Peter Hall, James Stephen Marron, and Amnon Neeman. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):427–444, 2005.

[Hon18] Luc M Hondeghem. Drug-induced qt prolongation and torsades de pointes: an all-exclusive relationship or time for an amicable separation? *Drug safety*, 41(1):11–17, 2018.

[HQ06] Richard A Helms and David J Quan. *Textbook of therapeutics: drug and disease management*. Lippincott Williams & Wilkins, 2006.

[HTC+14] Michael Hay, David W Thomas, John L Craighead, Celia Economides, and Jesse Rosenthal. Clinical development success rates for investigational drugs. *Nature biotechnology*, 32(1):40–51, 2014.

[HTW19] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2019.

[Hug68] Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.

[HVWR11] Jordi Heijman, Paul GA Volders, Ronald L Westra, and Yoram Rudy. Local control of β-adrenergic stimulation: effects on ventricular myocyte electrophysiology and ca2+-transient. *Journal of molecular and cellular cardiology*, 50(5):863–871, 2011.

[HZJ+13] Jordi Heijman, Antonio Zaza, Daniel M Johnson, Yoram Rudy, Ralf LM Peeters, Paul GA Volders, and Ronald L Westra. Determinants of beat-to-beat variability of repolarization duration in the canine ventricular myocyte: a computational analysis. *PLoS computational biology*, 9(8):e1003202, 2013.

[IB09] M Willjuice Iruthayarajan and S Baskar. Evolutionary algorithms based design of multivariable pid controller. *Expert Systems with applications*, 36(5):9159–9167, 2009.

[INNW+17] Hiroko Izumi-Nakaseko, Yuji Nakamura, Takeshi Wada, Kentaro Ando, Yasunari Kanda, Yuko Sekino, and Atsushi Sugiyama. Characterization of human ips cell-derived cardiomyocyte sheets as a model to detect drug-induced conduction disturbance. *The Journal of toxicological sciences*, 42(2):183–192, 2017.

[JBG+16] Ross H Johnstone, Rémi Bardenet, David J Gavaghan, Liudmila Polonchuk, Mark R Davies, and Gary R Mirams. Hierarchical bayesian modelling of variability and uncertainty in synthetic action potential traces. In *2016 Computing in Cardiology Conference (CinC)*, pages 1089–1092. IEEE, 2016.

[JCB+16] Ross H Johnstone, Eugene TY Chang, Rémi Bardenet, Teun P De Boer, David J Gavaghan, Pras Pathmanathan, Richard H Clayton, and Gary R Mirams. Uncertainty and variability in models of the cardiac action potential: Can we build trustworthy models? *Journal of molecular and cellular cardiology*, 96:49–62, 2016.

[JCW+20] Karoline Horgmo Jæger, Verena Charwat, Samuel Wall, Kevin E Healy, and Aslak Tveito. Identifying drug response by combining measurements of the membrane potential, the cytosolic calcium concentration, and the extracellular potential in microphysiological systems. *Frontiers in pharmacology*, 11, 2020.

[JHB+13] Daniel M Johnson, Jordi Heijman, Elizabeth F Bode, David J Greensmith, Henk van der Linde, Najah Abi-Gerges, David A Eisner, Andrew W Trafford, and Paul GA Volders. Diastolic spontaneous calcium release from the sarcoplasmic reticulum increases beat-to-beat variability of repolarization in canine ventricular myocytes after β-adrenergic stimulation. *Circulation research*, 112(2):246–256, 2013.

[JHP+10] Daniel M Johnson, Jordi Heijman, Chris E Pollard, Jean-Pierre Valentin, Harry JGM Crijns, Najah Abi-Gerges, and Paul GA Volders. Iks restricts excessive beat-to-beat variability of repolarization during beta-adrenergic receptor stimulation. *Journal of molecular and cellular cardiology*, 48(1):122–130, 2010.

[JKT+03] Yasuhiko Jimbo, Nahoko Kasai, Keiichi Torimitsu, Takashi Tateno, and Hugh PC Robinson. A system for mea-based multisite stimulation. *IEEE transactions on biomedical engineering*, 50(2):241–248, 2003.

[Joh03] Mikael Johnson. Classification of ae transients based on numerical simulations of composite laminates. *Ndt & e International*, 36(5):319–329, 2003.

[JU04] Simon J Julier and Jeffrey K Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004.

[JVB+05] Norbert Jost, László Virág, Miklós Bitay, János Takács, Csaba Lengyel, Péter Biliczki, Zsolt Nagy, Gábor Bogáts, David A Lathrop, Julius G Papp, et al. Restricting excessive cardiac action potential and qt prolongation: a vital role for i ks in human ventricular muscle. *Circulation*, 112(10):1392–1399, 2005.

[JVD+10] MKB Jonsson, MA Vos, G Duker, S Demolombe, and TAB Van Veen. Gender disparity in cardiac electrophysiology: implications for cardiac safety pharmacology. *Pharmacology & therapeutics*, 127(1):9–18, 2010.

[JVM+12] Malin KB Jonsson, Marc A Vos, Gary R Mirams, Göran Duker, Peter Sartipy, Teun P De Boer, and Toon AB Van Veen. Application of human stem cell-derived cardiomyocytes in safety pharmacology requires caution beyond herg. *Journal of molecular and cellular cardiology*, 52(5):998–1008, 2012.

[JVS+08] Norbert Jost, Andras Varro, Viktoria Szuts, Peter P Kovacs, Gyorgy Seprenyi, Peter Biliczki, Csaba Lengyel, Janos Prorok, Miklos Bitay, Balazs Ordog, et al. Molecular basis of repolarization reserve differences between dogs and man, 2008.

[KATW15] Ioannis Karakikes, Mohamed Ameen, Vittavat Termglinchan, and Joseph C Wu. Human induced pluripotent stem cell–derived cardiomyocytes: insights into molecular, cellular, and functional phenotypes. *Circulation research*, 117(1):80–88, 2015.

[KC13] Jacek Z. Kubiak and Maria A. Ciemerych. From Gurdon to Yamanaka, short story of cell reprogramming / Od Gurdona do Yamanaki, czyli krytka historia reprogramowania komyrek. *Postepy biochemii*, 59(2):124–30, January 2013.

[KG96] Nalin M Kumar and Norton B Gilula. The gap junction communication channel. *Cell*, 84(3):381–388, 1996.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[Kla11] Richard Klabunde. *Cardiovascular physiology concepts*. Lippincott Williams & Wilkins, 2011.

[KMH+04] Stefan Kern, Sibylle D Müller, Nikolaus Hansen, Dirk Büche, Jiri Ocenasek, and Petros Koumoutsakos. Learning probability distributions in continuous evolutionary algorithms–a comparative review. *Natural Computing*, 3(1):77–112, 2004.

[KNG+19] Lucas Karperien, Ali Navaei, Brent Godau, Alireza Dolatshahi-Pirouz, Mohsen Akbari, and Mehdi Nikkhah. Nanoengineered biomaterials for cardiac regeneration. In *Nanoengineered Biomaterials for Regenerative Medicine*, pages 95–124. Elsevier, 2019.

[KOPM+13] James Kramer, Carlos A Obejero-Paz, Glenn Myatt, Yuri A Kuryshev, Andrew Bruening-Wright, Joseph S Verducci, and Arthur M Brown. Mice models: superior to the herg model in predicting torsade de pointes. *Scientific reports*, 3(1):1–7, 2013.

[Kor07] Bruce G Kornreich. The patch clamp technique: principles and technical considerations. *Journal of Veterinary Cardiology*, 9(1):25–37, 2007.

[KPCK20] Mikhail Kushnarev, Iulia Paula Pirvulescu, Kenneth D Candido, and Nebojsa Nick Knezevic. Neuropathic pain: preclinical and early clinical progress with voltage-gated sodium channel blockers. *Expert opinion on investigational drugs*, 29(3):259–271, 2020.

[KPZ07] Hyunsoo Kim, Haesun Park, and Hongyuan Zha. Distance preserving dimension reduction for manifold learning. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 527–532. SIAM, 2007.

[KS01] Shigeto Kanno and Jeffrey E Saffitz. The role of myocardial gap junctions in electrical conduction and arrhythmogenesis. *Cardiovascular pathology*, 10(4):169–177, 2001.

[KSH+15] K Kistamas, N Szentandrassy, B Hegyi, K Vaczi, F Ruzsnavszky, B Horvath, T Banyasz, PP Nanasi, and J Magyar. Changes in intracellular calcium concentration influence beat-to-beat variability of action potential duration in canine ventricular myocytes. *J. Physiol. Pharmacol*, 66(1):73–81, 2015.

[Küg20] Philipp Kügler. Modelling and simulation for preclinical cardiac safety assessment of drugs with human ipsc-derived cardiomyocytes. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 122(4):209–257, 2020.

[KV02] Lewis B Kinter and Jean-Pierre Valentin. Safety pharmacology and risk assessment. *Fundamental & clinical pharmacology*, 16(3):175–182, 2002.

[KYO+18] Yasunari Kanda, Daiju Yamazaki, Tomoharu Osada, Takashi Yoshinaga, and Kohei Sawada. Development of torsadogenic risk assessment using human induced pluripotent stem cell-derived cardiomyocytes: Japan ips cardiac safety assessment (jicsa) update. *Journal of pharmacological sciences*, 138(4):233–239, 2018.

[LBF11] Yi Liu, Edward J Beck, and Christopher M Flores. Validation of a patch clamp screening protocol that simultaneously measures compound activity in multiple states of the voltage-gated sodium channel nav1. 2. *Assay and drug development technologies*, 9(6):628–634, 2011.

[LBM+17] Hendrik Lapp, Tobias Bruegmann, Daniela Malan, Stephanie Friedrichs, Carsten Kilgus, Alexandra Heidsieck, and Philipp Sasse. Frequency-dependent drug screening using optogenetic stimulation of human ipsc-derived cardiomyocytes. *Scientific reports*, 7(1):1–12, 2017.

[LDC+18] Brodie AJ Lawson, Christopher C Drovandi, Nicole Cusimano, Pamela Burrage, Blanca Rodriguez, and Kevin Burrage. Unlocking data sets by calibrating populations of models to data density: A study in atrial electrophysiology. *Science advances*, 4(1):e1701676, 2018.

[LdLK11] Mathieu Lemay, Enno de Lange, and Jan P Kucera. Effects of stochastic channel gating and distribution on the cardiac action potential. *Journal of theoretical biology*, 281(1):84–96, 2011.

[Lei20] Derek J Leishman. Improving prediction of torsadogenic risk in the cipa in silico model by appropriately accounting for clinical exposure. *Journal of pharmacological and toxicological methods*, 101:106654, 2020.

[LFGS16] Maïlys Lopes, Mathieu Fauvel, Stéphane Girard, and David Sheeren. High dimensional kullback-leibler divergence for grassland management practices classification from high resolution satellite image time series. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3342–3345. IEEE, 2016.

[LFYC98] Gui-Rong Li, Jianlin Feng, Lixia Yue, and Michel Carrier. Transmural heterogeneity of action potentials and i to1 in myocytes isolated from the human right ventricle. *American Journal of Physiology-Heart and Circulatory Physiology*, 275(2):H369–H377, 1998.

[LHB+18] Hyang-Ae Lee, Sung-Ae Hyun, Byungjin Byun, Jong-Hak Chae, and Ki-Suk Kim. Electrophysiological mechanisms of vandetanib-induced cardiotoxicity: comparison of action potentials in rabbit purkinje fibers and pluripotent stem cell-derived cardiomyocytes. *PloS one*, 13(4):e0195577, 2018.

[LHNM08] Yufeng Liu, David Neil Hayes, Andrew Nobel, and James Stephen Marron. Statistical significance of clustering for high-dimension, low–sample size data. *Journal of the American Statistical Association*, 103(483):1281–1293, 2008.

[LHS14] Maksim Lapin, Matthias Hein, and Bernt Schiele. Learning using privileged information: Svm+ and weighted svm. *Neural Networks*, 53:95–108, May 2014.

[Lin13] John E Linley. Perforated whole-cell patch-clamp recording. In *Ion Channels*, pages 149–157. Springer, 2013.

[LKK+06] So-Young Lee, Young-Jin Kim, Kyong-Tai Kim, Han Choe, and Su-Hyun Jo. Blockade of herg human k+ channels and ikr of guinea-pig cardiomyocytes by the antipsychotic drug clozapine. *British journal of pharmacology*, 148(4):499–509, 2006.

[LKMGG07] Claudia Lerma, Trine Krogh-Madsen, Michael Guevara, and Leon Glass. Stochastic aspects of cardiac arrhythmias. *Journal of Statistical Physics*, 128(1):347–374, 2007.

[LMY+20] Zhihua Li, Gary R Mirams, Takashi Yoshinaga, Bradley J Ridder, Xiaomei Han, Janell E Chen, Norman L Stockbridge, Todd A Wisialowski, Bruce Damiano, Stefano Severi, et al. General principles for the validation of proarrhythmia risk prediction models: an extension of the cipa in silico strategy. *Clinical Pharmacology & Therapeutics*, 107(1):102–111, 2020.

[LR19] Damiano Lombardi and Fabien Raphel. A greedy dimension reduction method for classification problems. 2019.

[LR21] Damiano Lombardi and Fabien Raphel. A method to enrich experimental datasets by means of numerical simulations in view of classification tasks. 2021.

[LRH+19] Zhihua Li, Bradley J Ridder, Xiaomei Han, Wendy W Wu, Jiansong Sheng, Phu N Tran, Min Wu, Aaron Randolph, Ross H Johnstone, Gary R Mirams, et al. Assessment of an in silico mechanistic model for proarrhythmia risk prediction under the ci pa initiative. *Clinical Pharmacology & Therapeutics*, 105(2):466–475, 2019.

[LS03] Ce Liu and Hueng-Yeung Shum. Kullback-leibler boosting. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[LS16] M Cummins Lancaster and EA Sobie. Improved prediction of drug-induced torsades de pointes through simulations of dynamics and machine learning algorithms. *Clinical Pharmacology & Therapeutics*, 100(4):371–379, 2016.

[LSP15] Sarah D Lamore, Clay W Scott, and Matthew F Peters. Cardiomyocyte impedance assays. 2015.

[LT83] KS Lee and RW Tsien. Mechanism of calcium channel blockade by verapamil, d600, diltiazem and nitrendipine in single dialysed heart cells. *Nature*, 302(5911):790–794, 1983.

[LW19] Nicolas Langrené and Xavier Warin. Fast and stable multivariate kernel density estimation by fast sum updating. *Journal of Computational and Graphical Statistics*, 28(3):596–608, 2019.

[LWZY17] Bo Liu, Ying Wei, Yu Zhang, and Qiang Yang. Deep neural networks for high dimension, low sample size data. In *IJCAI*, pages 2287–2293, 2017.

[LXEB+20] Huan Lan, Qiang Xu, Ibrahim El-Battrawy, Rujia Zhong, Xin Li, Siegfried Lang, Lukas Cyganek, Martin Borggrefe, Xiaobo Zhou, and Ibrahim Akin. Ionic mechanisms of disopyramide prolonging action potential duration in human-induced pluripotent stem cell-derived cardiomyocytes from a patient with short qt syndrome type 1. *Frontiers in pharmacology*, 11:1613, 2020.

[Mao15] Xuerong Mao. The truncated euler–maruyama method for stochastic differential equations. *Journal of Computational and Applied Mathematics*, 290:370–384, 2015.

[Mar08] Elena Marchiori. Hit miss networks with applications to instance selection. 2008.

[MBGF04] Thomas Meyer, Karl-Heinz Boven, Elke Günther, and Michael Fejtl. Microelectrode arrays in cardiac safety pharmacology. *Drug Safety*, 27(11):763–772, 2004.

[MCS+11] Gary R Mirams, Yi Cui, Anna Sher, Martin Fink, Jonathan Cooper, Bronagh M Heath, Nick C McMahon, David J Gavaghan, and Denis Noble. Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk. *Cardiovascular research*, 91(1):53–61, 2011.

[MdKD+18] Petra Mulder, Tessa de Korte, Elena Dragicevic, Udo Kraushaar, Richard Printemps, Maria LH Vlaming, Stefan R Braam, and Jean-Pierre Valentin. Predicting cardiac safety using human induced pluripotent stem cell-derived cardiomyocytes combined with multi-electrode array (mea) technology: a conference report. *Journal of pharmacological and toxicological methods*, 91:36–42, 2018.

[MDS+18] Daniel Millard, Qianyu Dang, Hong Shi, Xiaou Zhang, Chris Strock, Udo Kraushaar, Haoyu Zeng, Paul Levesque, Hua-Rong Lu, Jean-Michel Guillon, et al. Cross-site reliability of human induced pluripotent stem cell-derived cardiomyocyte based safety assays using microelectrode arrays: results from a blinded cipa pilot study. *Toxicological Sciences*, 164(2):550–562, 2018.

[Mec12] Elizabeth Meckes. Approximation of projections of random vectors. *Journal of Theoretical Probability*, 25(2):333–352, 2012.

[MFH+18] Andrea Mendizabal, Tatiana Fountoukidou, Jan Hermann, Raphael Sznitman, and Stéphane Cotin. A combined simulation and machine learning approach for image-based force classification during robotized intravitreal injections. In *International conference on medical image computing and computer-assisted intervention*, pages 12–20. Springer, 2018.

[MGF+11] Junyi Ma, Liang Guo, Steve J Fiene, Blake D Anson, James A Thomson, Timothy J Kamp, Kyle L Kolaja, Bradley J Swanson, and Craig T January. High purity human-induced pluripotent stem cell-derived cardiomyocytes: electrophysiological properties of action potentials and ionic currents. *American Journal of Physiology-Heart and Circulatory Physiology*, 301(5):H2006–H2017, 2011.

[MHB96] B Müller, A Hasman, and JA Blom. Building intelligent alarm systems by combining mathematical models and inductive machine learning techniques part 2—sensitivity analysis. *International journal of bio-medical computing*, 42(3):165–179, 1996.

[MSMSR05] MJ Mason, AK Simpson, MP Mahaut-Smith, and HPC Robinson. The interpretation of current-clamp recordings in the cell-attached patch-clamp configuration. *Biophysical journal*, 88(1):739–750, 2005.

[MUV15] Sebastian Mayer, Tino Ullrich, and Jan Vybiral. Entropy and sampling numbers of classes of ridge functions. *Constructive Approximation*, 42(2):231–264, 2015.

[MWZ+13] Dongrui Ma, Heming Wei, Yongxing Zhao, Jun Lu, Guang Li, Norliza Binte Esmail Sahib, Teng Hong Tan, Keng Yean Wong, Winston Shim, Philip Wong, et al. Modeling type 3 long qt syndrome with cardiomyocytes derived from patient-specific induced pluripotent stem cells. *International journal of cardiology*, 168(6):5277–5286, 2013.

[Nar05] Anand Narasimhamurthy. Theoretical bounds of majority voting performance for a binary classification problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1988–1995, 2005.

[Nav13] Jorge Navarro. A simple proof for the multivariate chebyshev inequality. *arXiv preprint arXiv:1305.5646*, 2013.

[NAY62] Jinichi Nagumo, Suguru Arimoto, and Shuji Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.

[NE14] David Nova and Pablo A Estévez. A review of learning vector quantization classifiers. *Neural Computing and Applications*, 25(3):511–524, 2014.

[NHW+17] Yumiko Nozaki, Yayoi Honda, Hitoshi Watanabe, Shota Saiki, Kiyotaka Koyabu, Tetsuji Itoh, Chiho Nagasawa, Chiaki Nakamori, Chiaki Nakayama, Hiroshi Iwasaki, et al. Csahi study-2: validation of multi-electrode array systems (mea60/2100) for prediction of drug-induced proarrhythmia using human ips cell-derived cardiomyocytes: assessment of reference compounds and comparison with non-clinical studies and clinical information. *Regulatory Toxicology and Pharmacology*, 88:238–251, 2017.

[NML+15] Jorge A Negroni, Stefano Morotti, Elena C Lascano, Aldrin V Gomes, Eleonora Grandi, José L Puglisi, and Donald M Bers. β-adrenergic effects on cardiac myofilaments and contraction in an integrated rabbit ventricular myocyte model. *Journal of molecular and cellular cardiology*, 81:162–175, 2015.

[NMM+14] Yuji Nakamura, Junko Matsuo, Norimasa Miyamoto, Atsuko Ojima, Kentaro Ando, Yasunari Kanda, Kohei Sawada, Atsushi Sugiyama, and Yuko Sekino. Assessment of testing methods for drug-induced repolarization delay and arrhythmias in an ips cell–derived cardiomyocyte sheet: Multi-site validation study. *Journal of pharmacological sciences*, page 13248FP, 2014.

[NP16] Ivan Nourdin and Guillaume Poly. Convergence in law implies convergence in total variation for polynomials in independent gaussian, gamma or beta random variables. In *High Dimensional Probability VII*, pages 381–394. Springer, 2016.

[OJT+16] Alison Obergrussberger, Krisztina Juhasz, Ulrich Thomas, Sonja Stölzle-Feix, Nadine Becker, Leo Dörr, Matthias Beckler, Corina Bot, Michael George, and Niels Fertig. Safety pharmacology studies using efp and impedance. *Journal of pharmacological and toxicological methods*, 81:223–232, 2016.

[ONN89] NOBUKUNI Ogata, MASAO Nishimura, and TOSHIO Narahashi. Kinetics of chlorpromazine block of sodium channels in single guinea pig cardiac myocytes. *Journal of Pharmacology and Experimental Therapeutics*, 248(2):605–613, 1989.

[OR12] Thomas O'Hara and Yoram Rudy. Quantitative comparison of cardiac ventricular myocyte electrophysiology and response to drugs in human and nonhuman species. *American Journal of Physiology-Heart and Circulatory Physiology*, 302(5):H1023–H1030, 2012.

[OVVR11] Thomas O'Hara, László Virág, András Varró, and Yoram Rudy. Simulation of the undiseased human cardiac ventricular action potential: model formulation and experimental validation. *PLoS computational biology*, 7(5):e1002061, 2011.

[PAC08] Michael K Pugsley, S Authier, and MJ Curtis. Principles of safety pharmacology. *British journal of pharmacology*, 154(7):1382–1399, 2008.

[PBC05] Andrew Pullan, Martin L Buist, and Leo K Cheng. *Mathematically modelling the electrical activity of the heart: from cell to body surface and back again*. World Scientific Publishing Company, 2005.

[PBL+17] Elisa Passini, Oliver J Britton, Hua Rong Lu, Jutta Rohrbacher, An N Hermans, David J Gallacher, Robert JH Greig, Alfonso Bueno-Orovio, and Blanca Rodriguez. Human in silico drug trials demonstrate higher accuracy than animal models in predicting clinical pro-arrhythmic cardiotoxicity. *Frontiers in physiology*, 8:668, 2017.

[PCV+11] Esther Pueyo, Alberto Corrias, László Virág, Norbert Jost, Tamás Szél, András Varró, Norbert Szentandrássy, Péter P Nánási, Kevin Burrage, and Blanca Rodríguez. A multiscale investigation of repolarization variability and its role in cardiac arrhythmogenesis. *Biophysical journal*, 101(12):2892–2902, 2011.

[PDB+18] E Pueyo, CE Dangerfield, OJ Britton, L Virág, K Kistamás, N Szentandrássy, N Jost, A Varró, PP Nánási, K Burrage, et al. Correction: Experimentally-based computational investigation into beat-to-beat variability in ventricular repolarization and its response to ionic current inhibition. *Plos one*, 13(5):e0197871, 2018.

[PGLC15] Vishal M Patel, Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE signal processing magazine*, 32(3):53–69, 2015.

[PHASS13] Michelangelo Paci, Jari Hyttinen, Katriina Aalto-Setälä, and Stefano Severi. Computational models of ventricular-and atrial-like human induced pluripotent stem cell derived cardiomyocytes. *Annals of biomedical engineering*, 41(11):2334–2348, 2013.

[PHdK+18] Michael K Pugsley, Marci L Harter, Tessa de Korte, Chantelle Connaughton, Simon Authier, and Michael J Curtis. Safety pharmacology methods and regulatory considerations evolve together. *J Pharmacol Toxicol Methods*, pages 1–6, 2018.

[Phi06] Jean Philibert. One and a half century of diffusion: Fick, einstein before and beyond. 2006.

[Pin06] Jerome Pine. A history of mea development. In *Advances in network electrophysiology*, pages 3–23. Springer, 2006.

[PJYK19] Jin-Sol Park, Ji-Young Jeon, Ji-Ho Yang, and Min-Gul Kim. Introduction to in silico model for proarrhythmic risk assessment under the cipa initiative. *Translational and clinical pharmacology*, 27(1):12–18, 2019.

[PMPI+02] Burkert Pieske, Lars S Maier, Valentino Piacentino III, Jutta Weisser, Gerd Hasenfuss, and Steven Houser. Rate dependence of [na+] i and contractility in nonfailing and failing human myocardium. *Circulation*, 106(4):447–453, 2002.

[PORT16] Esther Pueyo, Michele Orini, José F Rodríguez, and Peter Taggart. Interactive effect of beta-adrenergic stimulation and mechanical stretch on low-frequency oscillations of ventricular action potential duration in humans. *Journal of molecular and cellular cardiology*, 97:93–105, 2016.

[PPC+18] Michelangelo Paci, Risto-Pekka Pölönen, Dario Cori, Kirsi Penttinen, Katriina Aalto-Setälä, Stefano Severi, and Jari Hyttinen. Automatic optimization of an in silico model of human ipsc derived cardiomyocytes recapitulating calcium handling abnormalities. *Frontiers in physiology*, 9:709, 2018.

[Pre98] Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, 1998.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[QSHZ15] Chen Qin, Shiji Song, Gao Huang, and Lei Zhu. Unsupervised neighborhood component analysis for clustering. *Neurocomputing*, 168:609–617, 2015.

[QX18] Bowen Qin and Fuyuan Xiao. A non-parametric method to determine basic probability assignment based on kernel density estimation. *IEEE Access*, 6:73509–73519, 2018.

[RBZ+17] Fabien Raphel, Muriel Boulakia, Nejib Zemzemi, Yves Coudière, Jean-Michel Guillon, Philippe Zitoun, and Jean-Frédéric Gerbeau. Identification of ion currents components generating field potential recorded in mea from hipsc-cm. *IEEE Transactions on Biomedical Engineering*, 65(6):1311–1319, 2017.

[RCAG+15] Blanca Rodriguez, Annamaria Carusi, Najah Abi-Gerges, Rina Ariga, Oliver Britton, Gil Bub, Alfonso Bueno-Orovio, Rebecca AB Burton, Valentina Carapella, Louie Cardone-Noott, et al. Human-based approaches to pharmacology and cardiology: an interdisciplinary and intersectorial workshop. *Ep Europace*, 18(9):1287–1298, 2015.

[RDKL+20] Fabien Raphel, Tessa De Korte, Damiano Lombardi, Stefan Braam, and Jean-Frederic Gerbeau. A greedy classifier optimization strategy to assess ion channel blocking activity and pro-arrhythmia in hipsc-cardiomyocytes. *PLoS computational biology*, 16(9):e1008203, 2020.

[RK05] Roman Rosipal and Nicole Krämer. Overview and recent advances in partial least squares. In *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pages 34–51. Springer, 2005.

[RMMC20] Luca Rosafalco, Andrea Manzoni, Stefano Mariani, and Alberto Corigliano. Fully convolutional networks for structural health monitoring through multivariate time series classification. *Advanced Modeling and Simulation in Engineering Sciences*, 7(1):1–31, 2020.

[Rod14] Dan M Roden. Pharmacology and toxicology of nav1. 5-class 1 antiarrhythmic drugs. *Cardiac electrophysiology clinics*, 6(4):695–704, 2014.

[RPM+05] Maurizio Recanatini, Elisabetta Poluzzi, Matteo Masetti, Andrea Cavalli, and Fabrizio De Ponti. Qt prolongation through herg k+ channel blockade: current knowledge and strategies for the early prediction during drug development. *Medicinal research reviews*, 25(2):133–166, 2005.

[RSB+04] Javier Ramírez, Jaume C Segura, Carmen Benítez, Angel De La Torre, and Antonio J Rubio. A new kullback-leibler vad for speech recognition in noise. *IEEE signal processing letters*, 11(2):266–269, 2004.

[RWLI75] G Ritter, H Woodruff, S Lowry, and T Isenhour. An algorithm for a selective nearest neighbor decision rule (corresp.). *IEEE Transactions on Information Theory*, 21(6):665–669, 1975.

[RWP+02] William S Redfern, Ian D Wakefield, Helen Prior, Christopher E Pollard, Timothy G Hammond, and Jean-Pierre Valentin. Safety pharmacology–a progressive approach. *Fundamental & Clinical Pharmacology*, 16(3):161–173, 2002.

[Sär13] Simo Särkkä. *Bayesian filtering and smoothing*. Number 3. Cambridge University Press, 2013.

[SBM17] Luca Sala, Milena Bellin, and Christine L Mummery. Integrating cardiomyocytes from human pluripotent stem cells in safety pharmacology: has the time come? *British journal of pharmacology*, 174(21):3749–3765, 2017.

[SBOW+14] Carlos Sánchez, Alfonso Bueno-Orovio, Erich Wettwer, Simone Loose, Jana Simon, Ursula Ravens, Esther Pueyo, and Blanca Rodriguez. Inter-subject variability in human atrial action potential in sinus rhythm versus chronic atrial fibrillation. *PloS one*, 9(8):e105897, 2014.

[SCG16] Elisa Schenone, Annabelle Collin, and Jean-Frédéric Gerbeau. Numerical simulation of electrocardiograms for full cardiac cycles in healthy and pathological conditions. *International journal for numerical methods in biomedical engineering*, 32(5):e02744, 2016.

[SCS12] Amrita X Sarkar, David J Christini, and Eric A Sobie. Exploiting mathematical models to illuminate electrophysiological variability between individuals. *The Journal of physiology*, 590(11):2555–2567, 2012.

[SEG+03] Alfred Stett, Ulrich Egert, Elke Guenther, Frank Hofmann, Thomas Meyer, Wilfried Nisch, and Hugo Haemmerle. Biological application of microelectrode arrays in drug discovery and basic research. *Analytical and bioanalytical chemistry*, 377(3):486–495, 2003.

[She14] Arkady Shemyakin. Hellinger distance and non-informative priors. *Bayesian Analysis*, 9(4):923–938, 2014.

[SHH+98] Ulrich Schmidt, Roger J Hajjar, Patrick A Helm, Catherine S Kim, Angelia A Doye, and Judith K Gwathmey. Contribution of abnormal sarcoplasmic reticulum atpase activity to systolic and diastolic dysfunction in human heart failure. *Journal of molecular and cellular cardiology*, 30(10):1929–1937, 1998.

[SIL07] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.

[SKH+15] Norbert Szentandrássy, Kornél Kistamás, Bence Hegyi, Balázs Horváth, Ferenc Ruzsnavszky, Krisztina Váczi, János Magyar, Tamás Bányász, András Varró, and Péter P Nánási. Contribution of ion currents to beat-to-beat variability of action potential duration in canine ventricular myocytes. *Pflügers Archiv-European Journal of Physiology*, 467(7):1431–1443, 2015.

[SN84] Bert Sakmann and Erwin Neher. Patch clamp techniques for studying ionic channels in excitable membranes. *Annual review of physiology*, 46(1):455–472, 1984.

[SPFBP17] David Adolfo Sampedro-Puente, Jesus Fernandez-Bes, and Esther Pueyo. Differential responses to beta-adrenergic stimulation in the long-qt syndrome type 1: Characterization and mechanisms. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

[SPFBP+19] David Adolfo Sampedro-Puente, Jesus Fernandez-Bes, Bradley Porter, Stefan Van Duijvenboden, Peter Taggart, and Esther Pueyo. Mechanisms underlying interactions between low-frequency oscillations and beat-to-beat variability of celullar ventricular repolarization in response to sympathetic stimulation: implications for arrhythmogenesis. *Frontiers in physiology*, 10:916, 2019.

[SPFBS+20] David Adolfo Sampedro-Puente, Jesus Fernandez-Bes, Norbert Szentandrássy, Péter Nánási, Peter Taggart, and Esther Pueyo. Time course of low-frequency oscillatory behavior in human ventricular repolarization following enhanced sympathetic activity and relation to arrhythmogenesis. *Frontiers in physiology*, 10:1547, 2020.

[SPFBV+19] David Adolfo Sampedro-Puente, Jesus Fernandez-Bes, László Virág, András Varró, and Esther Pueyo. Data-driven identification of stochastic model parameters and state variables: Application to the study of cardiac beat-to-beat variability. *IEEE journal of biomedical and health informatics*, 24(3):693–704, 2019.

[SPRFB+20] David Adolfo Sampedro-Puente, Fabien Raphel, Jesus Fernandez-Bes, Pablo Laguna, Damiano Lombardi, and Esther Pueyo. Characterization of spatio-temporal cardiac action potential variability at baseline and under β-adrenergic stimulation by combined unscented kalman filter and double greedy dimension reduction. *IEEE journal of biomedical and health informatics*, 25(1):276–288, 2020.

[SS10] Anthony R Soltis and Jeffrey J Saucerman. Synergy between camkii substrates and β-adrenergic signaling in regulation of cardiac myocyte ca2+ handling. *Biophysical journal*, 99(7):2038–2047, 2010.

[SSLN10] Biswa Sengupta, Martin Stemmler, Simon B Laughlin, and Jeremy E Niven. Action potential energy efficiency varies among neuron types in vertebrates and invertebrates. *PLoS computational biology*, 6(7):e1000840, 2010.

[SSW15] Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92, 2015.

[Ste13] David C. Sterratt. *Nernst Equation*, pages 1–3. Springer New York, New York, NY, 2013.

[Sur21] Ananda Theertha Suresh. Robust hypothesis testing and distribution estimation in hellinger distance. In *International Conference on Artificial Intelligence and Statistics*, pages 2962–2970. PMLR, 2021.

[SVD+17] Carsten Stoetzer, Marc Voelker, Thorben Doll, Joerg Heineke, Florian Wegner, and Andreas Leffler. Cardiotoxic antiemetics metoclopramide and domperidone block cardiac voltage-gated na+ channels. *Anesthesia & Analgesia*, 124(1):52–60, 2017.

[SWY+11] Yoshihiro Sobue, Eiichi Watanabe, Mayumi Yamamoto, Kan Sano, Hiroto Harigaya, Kentarou Okuda, and Yukio Ozaki. Beat-to-beat variability of t-wave amplitude for the risk assessment of ventricular tachyarrhythmia in patients without structural heart disease. *Europace*, 13(11):1612–1618, 2011.

[SZ03] Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01):17–41, 2003.

[SZD+04] Gernot Schram, Liming Zhang, Katayoun Derakhchan, Joachim R Ehrlich, Luiz Belardinelli, and Stanley Nattel. Ranolazine: ion-channel-blocking actions and in vivo electrophysiological effects. *British journal of pharmacology*, 142(8):1300–1308, 2004.

[T+76] Ivan Tomek et al. An experiment with the edited nearest-nieghbor rule. 1976.

[Tad98] Saldju Tadjudin. *Classification of high dimensional data with limited training samples*. Purdue University, 1998.

[TBVM+18] LGJ Tertoolen, SR Braam, BJ Van Meer, R Passier, and CL Mummery. Interpretation of field potentials measured on a multi electrode array in pharmacological toxicity screening on primary and human pluripotent stem cell-derived cardiomyocytes. *Biochemical and biophysical research communications*, 497(4):1135–1141, 2018.

[Ted06] Luis Orlindo Tedeschi. Assessment of the adequacy of mathematical models. *Agricultural systems*, 89(2-3):225–247, 2006.

[Tem15] Vladimir Temlyakov. Sparse approximation by greedy algorithms. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 183–215. Springer, 2015.

[TF92] MB Tyers and AJ Freeman. Mechanism of the anti-emetic activity of 5-ht3 receptor antagonists. *Oncology*, 49(4):263–268, 1992.

[TGOW05] Antti J Tanskanen, Joseph L Greenstein, Brian O'Rourke, and Raimond L Winslow. The role of stochastic and modal gating of cardiac l-type ca2+ channels on early after-depolarizations. *Biophysical Journal*, 88(1):85–95, 2005.

[TJSL+72] CA Thomas Jr, PA Springer, GE Loeb, Y Berwald-Netter, and LM Okun. A miniature microelectrode array to monitor the bioelectric activity of cultured cells. *Experimental cell research*, 74(1):61–66, 1972.

[TLFR13] Athanasios Tsanas, Max A Little, Cynthia Fox, and Lorraine O Ramig. Objective automatic assessment of rehabilitative speech treatment in parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):181–190, 2013.

[TLRG17] Eliott Tixier, Damiano Lombardi, Blanca Rodriguez, and Jean-Frédéric Gerbeau. Modelling variability in cardiac electrophysiology: a moment-matching approach. *Journal of the Royal Society Interface*, 14(133):20170238, 2017.

[TPYP18] Tommaso Taddei, JD Penn, M Yano, and Anthony T Patera. Simulation-based classification; a model-order-reduction approach for structural health monitoring. *Archives of Computational Methods in Engineering*, 25(1):23–45, 2018.

[TRLG18] Eliott Tixier, Fabien Raphel, Damiano Lombardi, and Jean-Frédéric Gerbeau. Composite biomarkers derived from micro-electrode array measurements and computer simulations improve the classification of drug-induced channel block. *Frontiers in physiology*, 8:1096, 2018.

[Tru79] Gerard V Trunk. A problem of dimensionality: A simple example. *IEEE Transactions on pattern analysis and machine intelligence*, (3):306–307, 1979.

[TSC+18] Nicholas Tarabelloni, Elisa Schenone, Annabelle Collin, Francesca Ieva, Anna Maria Paganoni, and Jean-Frederic Gerbeau. Statistical assessment and calibration of numerical ecg models. *JP Journal of Biostatistics*, 15(2):151–173, 2018.

[TTM+09] Tomofumi Tanaka, Shugo Tohyama, Mitsushige Murata, Fumimasa Nomura, Tomoyuki Kaneko, Hao Chen, Fumiyuki Hattori, Toru Egashira, Tomohisa Seki, Yohei Ohno, et al. In vitro pharmacologic testing using human induced pluripotent stem cell-derived cardiomyocytes. *Biochemical and biophysical research communications*, 385(4):497–502, 2009.

[TTO+07] Kazutoshi Takahashi, Koji Tanabe, Mari Ohnuki, Megumi Narita, Tomoko Ichisaka, Kiichiro Tomoda, and Shinya Yamanaka. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *cell*, 131(5):861–872, 2007.

[Tun78a] Leslie Tung. *A bi-domain model for describing ischemic myocardial dc potentials.* PhD thesis, Massachusetts Institute of Technology, 1978.

[Tun78b] Leslie Tung. *A bi-domain model for describing ischemic myocardial dc potentials.* PhD thesis, Massachusetts Institute of Technology, 1978.

[VSP05] Fernando Vázquez, J Salvador Sánchez, and Filiberto Pla. A stochastic approach to wilson's editing algorithm. In *Iberian conference on pattern recognition and image analysis*, pages 35–42. Springer, 2005.

[VWvdH+12] Rosanne Varkevisser, Sofieke C Wijers, Marcel AG van der Heyden, Jet DM Beekman, Mathias Meine, and Marc A Vos. Beat-to-beat variability of repolarization as a new biomarker for proarrhythmia in vivo. *Heart Rhythm*, 9(10):1718–1726, 2012.

[VZJ+18] Jose Vicente, Robbert Zusterzeel, Lars Johannesen, Jay Mason, Philip Sager, Vikram Patel, Murali K Matta, Zhihua Li, Jiang Liu, Christine Garnett, et al. Mechanistic model-informed proarrhythmic risk assessment of drugs: review of the "cipa" initiative and design of a prospective clinical validation study. *Clinical Pharmacology & Therapeutics*, 103(1):54–66, 2018.

[WBD+18] Rob Wallis, Charles Benson, Borje Darpo, Gary Gintant, Yasunari Kanda, Krishna Prasad, David G Strauss, and Jean-Pierre Valentin. Cipa challenges and opportunities from a non-clinical, clinical and regulatory perspectives. an overview of the safety pharmacology scientific discussion. *Journal of pharmacological and toxicological methods*, 93:15–25, 2018.

[WD18] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[WEG87] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[Wei97] James N Weiss. The hill equation revisited: uses and misuses. *The FASEB Journal*, 11(11):835–841, 1997.

[WGK+10] James N Weiss, Alan Garfinkel, Hrayr S Karagueuzian, Peng-Sheng Chen, and Zhilin Qu. Early afterdepolarizations and cardiac arrhythmias. *Heart rhythm*, 7(12):1891–1899, 2010.

[WM97] D Randall Wilson and Tony R Martinez. Instance pruning techniques. In *ICML*, volume 97, pages 400–411, 1997.

[WM00] D Randall Wilson and Tony R Martinez. Reduction techniques for instance-based learning algorithms. *Machine learning*, 38(3):257–286, 2000.

[WM03] Guido de Wert and Christine Mummery. Human embryonic stem cells: research, ethics and policy. *Human reproduction*, 18(4):672–682, 2003.

[WSH+15] Jonathan W Waks, Elsayed Z Soliman, Charles A Henrikson, Nona Sotoodehnia, Lichy Han, Sunil K Agarwal, Dan E Arking, David S Siscovick, Scott D Solomon, Wendy S Post, et al. Beat-to-beat spatiotemporal variability in the t vector is associated with sudden cardiac death in participants without left ventricular hypertrophy: The atherosclerosis risk in communities (aric) study. *Journal of the American Heart Association*, 4(1):e001357, 2015.

[WSL19] Chi Heem Wong, Kien Wei Siah, and Andrew W Lo. Estimation of clinical trial success rates and related parameters. *Biostatistics*, 20(2):273–286, 2019.

[WVDMH01] Eric A Wan, Rudolph Van Der Merwe, and Simon Haykin. The unscented kalman filter. *Kalman filtering and neural networks*, 5(2007):221–280, 2001.

[WZ07] Fei Wang and Changshui Zhang. Feature extraction by maximizing the average neighborhood margin. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[XGP+13] Yuanfang Xie, Eleonora Grandi, Jose L Puglisi, Daisuke Sato, and Donald M Bers. β-adrenergic stimulation activates early afterdepolarizations transiently via kinetic mismatch of pka targets. *Journal of molecular and cellular cardiology*, 58:153–161, 2013.

[XNZ08] Shiming Xiang, Feiping Nie, and Changshui Zhang. Learning a mahalanobis distance metric for data clustering and classification. *Pattern recognition*, 41(12):3600–3612, 2008.

[XWL14] Jia-Min Xu, Ce-Qun Wang, and Long-Nian Lin. Multi-channel in vivo recording techniques: signal processing of action potentials and local field potentials. *Sheng li xue bao:[Acta physiologica Sinica]*, 66(3):349–357, 2014.

[Yam07] Shinya Yamanaka. Strategies and new developments in the generation of patient-specific pluripotent stem cells. *Cell stem cell*, 1(1):39–49, 2007.

[YB85] Atsuko Yatani and Arthur M Brown. The calcium channel blocker nitrendipine blocks sodium channels in neonatal rat cardiac myocytes. *Circulation research*, 56(6):868–875, 1985.

[YBS86] A Yatani, AM Brown, and A Schwartz. Bepridil block of cardiac calcium and sodium channels. *Journal of Pharmacology and Experimental Therapeutics*, 237(1):9–17, 1986.

[YBS12] Magdi Yacoub, Thomas Brand, and Jan Schlueter. Think small: The zebrafish as a novel disease model in qcrc. In *Qatar Foundation Annual Research Forum Volume 2012 Issue 1*, volume 2012, page BMP116. Hamad bin Khalifa University Press (HBKU Press), 2012.

[YKI+18] Daiju Yamazaki, Takashi Kitaguchi, Masakazu Ishimura, Tomohiko Taniguchi, Atsuhiro Yamanishi, Daisuke Saji, Etsushi Takahashi, Masao Oguchi, Yuta Moriyama, Sanae Maeda, et al. Proarrhythmia risk prediction using human induced pluripotent stem cell-derived cardiomyocytes. *Journal of pharmacological sciences*, 136(4):249–256, 2018.

[YKN21] Shingo Yamaguchi, Masayuki Kaneko, and Mamoru Narukawa. Approval success rates of drug candidates based on target, action, modality, application, and their combinations. *Clinical and Translational Science*, 2021.

[YT97] JIAN-AN YAO and GEA-NY TSENG. Azimilide (ne-10064) can prolong or shorten the action potential duration in canine ventricular myocytes: Dependence on blockade of k, ca, and na channels. *Journal of cardiovascular electrophysiology*, 8(2):184–198, 1997.

[YXZ12] Xiao Yajuan, Liang Xin, and Li Zhiyuan. A comparison of the performance and application differences between manual and automated patch-clamp techniques. *Current chemical genomics*, 6:87, 2012.

[ZBS+13] Nejib Zemzemi, Miguel O Bernabeu, Javier Saiz, Jonathan Cooper, Pras Pathmanathan, Gary R Mirams, Joe Pitt-Francis, and Blanca Rodriguez. Computational assessment of drug-induced effects on the electrocardiogram: from ion channel to body surface potentials. *British journal of pharmacology*, 168(3):718–733, 2013.

[ZdKN+19] Anne Zwartsen, Tessa de Korte, Peter Nacken, Dylan W de Lange, Remco HS Westerink, and Laura Hondebrink. Cardiotoxicity screening of illicit drugs and new psychoactive substances (nps) in human ipsc-derived cardiomyocytes using microelectrode array (mea) recordings. *Journal of molecular and cellular cardiology*, 136:102–112, 2019.

[ZRTH03] Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie. 1-norm support vector machines. In *Advances in neural information processing systems*, page None. Citeseer, 2003.

[ZSMW13] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827. PMLR, 2013.

[ZSS+17] Huanqi Zhu, Kelsey S Scharnhorst, Adam Z Stieg, James K Gimzewski, Itsunari Minami, Norio Nakatsuji, Haruko Nakano, and Atsushi Nakano. Two dimensional electrophysiological characterization of human pluripotent stem cell-derived cardiomyocyte system. *Scientific reports*, 7(1):1–9, 2017.

[ZWC+19] Haoyu Zeng, Jixin Wang, Holly Clouse, Armando Lagrutta, and Frederick Sannajust. Resolving the reversed rate effect of calcium channel blockers on human-induced pluripotent stem cell-derived cardiomyocytes and the impact on in vitro cardiac safety evaluation. *Toxicological Sciences*, 167(2):573–580, 2019.