Politecnico di Milano

SCHOOL OF INDUSTRIAL AND INFORMATION ENGINEERING

Master of Science - Biomedical Engineering

# Statistical Characterization of Continuous Wavelet Transform Derived Spiking Activity From In-Vitro Human Neural Networks Improves Classification of LRRK2-Mutated Cells

Supervisor

**Prof. Riccardo Barbieri**

Co-Supervisor

**Ing. Riccardo Levi**

Candidate

**Salvatore Castelbuono - 919674**

**Academic Year 2020-2021**

Salvatore Castelbuono: *Statistical Characterization of Continuous Wavelet Transform Derived Spiking Activity From In-Vitro Human Neural Networks Improves Classification of LRRK2-Mutated Cells* | Master of Science in Biomedical Engineering, Politecnico di Milano.

Politecnico di Milano:
www.polimi.it

School of Industrial and Information Engineering:
www.ingindinf.polimi.it

# Contents

# List of Figures

# List of Tables

# Abstract

Multi-Electrode Array (MEA) has become a widely used technology for assessing both the activity of in-vivo and in-vitro neural cells. In-vitro neural networks have been employed as a mean for studying how genetic mutations related to neurodegenerative disorders impact the electro-physiological behaviour of the cells. The scientific community has identified the G2019S mutation on the LRRK2 gene as a common cause of neurodegeneration in subjects affected by Parkinson's disease.

This work presents a statistical characterization of the spiking activity of in-vitro human neural networks carrying the G2019S as compared to a control population of healthy cells. The neurons have been cultured on a microfluidic chip integrated onto a customized MEA, and their activity has been recorded during a Baseline phase and after 24 hours from a stimulation through Kainic Acid.

The acquired signals are processed by a newly developed spike sorting approach based on spike detection algorithm that uses Continuous Wavelet Transform. Then, the DBSCAN clustering algorithm has been used to automatically sort the results. The obtained spike trains have been modeled using the Point Process framework, assessing the goodness-of-fit of known state-of-art stochastic distributions by using the Kolgomorov-Smirnov test. Thus, the Inverse Gaussian model has been chosen as the best model and the features of each spike train estimated with the Hamiltonian Monte Carlo Sampling algorithm.

Classification tasks have been performed to distinguish LRRK2-mutated and healthy cells based on the statistical parameters estimated. The results obtained show an improvement compared to previous works, with an accuracy of 96.15 reached for the Random Forest classifier for the binary discrimination. Moreover, the same trend was depicted in the multiclass classification for the data acquired after 24 hours from the stimulation, with an accuracy equal to 74.80. Finally, a greater firing activity has been observed in mutated cells compared to the Baseline stage, whereas a considerable drop after the neurotoxic insult was found.

# Sommario

Nelle scorse decadi, reti neurali *in-vivo* sono state studiate per valutare gli effetti di modifiche genetiche correlate a disordini neurologici sul corretto funzionamento bioelettrico delle cellule.

L'attività extracellulare può essere registrata tramite l'utilizzo di Micro Electrode Arrays, sui quali vengono posizionate le culture cellulari. Questi dispositivi sono composti da una serie di piccoli elettrodi di dimensioni nell'ordine di centinaia di $\mu$m, i quali sono in grado di acquisire con precisione l'attività elettrica di un neurone.

Inoltre, lo sviluppo di nuove tecniche per le culture cellulari *in-vitro* in seguito alla scoperta dei fattori di Yamanaka [2] ha permesso di ampliare le opportunità per lo studio delle mutazioni genetiche che colpiscono le cellule neuronali. In particolare, fibroblasti che presentano una specifica modifica nel DNA possono essere convertiti in cellule staminali indotte pluripotenti (iPSC), dalle quali si possono derivare le cellule neuronali. Integrando quindi la tecnologia dei MEA e lo studio di culture di reti neuronali *in-vitro*, patologie come il morbo di Parkinson (PD) possono essere studiate ad un livello cellulare.

Diversi studi[3][4] hanno dimostrato come la mutazione G2019S del gene LRRK2 può essere riscontrata in pazienti afetti da PD. Il ruolo di tale mutazione però non è ad oggi ancora approfonditamente delineato. Per cui lo studio di cellule neuronali che presentano tale modifca del codice genetico può migliore la conoscenza della patologia.

I segnali acquisiti però necessitano di essere processati affinché si possano estrarre informazioni utili per delineare l'attività neuronale. A tal proposito, possono essere impiegati degli algoritmi di spike sorting per identificare e caratterizzare i neuroni sulla base dell'attività bioelettrica.

Il lavoro di questa tesi si pone lo scopo di sviluppare un algoritmo di spike sorting basato sulla Continuous Wavelet Transform per estrarre le informazioni atte a caratterizzare cellulle aventi una mutazione del gene LRRK2 e cellule sane. Inoltre, il modello di Point Process viene impiegato per descrivere statisticamente il comportamento delle cellule neuronali. In ultima istanza, una classificazione unsupervised è stata effettuata per discernere neuroni sani da neuroni mutati utilizzando le features estratte.

Il lavoro qui presentato si pone l'obiettivo di migliorare una prima analisi effettuata sul dataset presso il Politecnico di Milano in collaborazione con il Dipartimento di Neuromedicina e delle Scienze del Movimento della Norwegian University of Science and Techonology (NTNU) di Trondheim (Norvegia).

### Materiale e Metodi

Reti neurali sono state sviluppate *in-vitro* a partire da cellule staminali pluripotenti indotte derivanti da fibroblasti umani per mezzo dei fattori di Yamanaka.[2] Due gruppi di reti sono stati ottenuti, rispettivamente formati da sei rete neurali presentanti la mutazione G2019S del gene LRRK2 e sei reti sane utilizzate come popolazione di controllo. Le reti neurali sono state coltivate su un chip microfluidico composto da tre camere interconnesse e posizionato su un MEA personalizzato [1] (figura 1) avente 60 elettrodi. L'attività neuronale delle reti è stata registrata durante una fase di *Baseline* e dopo 24 ore da una stimolazione elettrochimica per mezzo dell'acido kainico.

I segnali sono stato quindi analizzati per mezzo di un nuovo algoritmo di spike sorting sviluppato. In prima istanza, un filtro digitale passabanda di tipo Butterworth dell'ottavo ordine è stato impiegato per filtrare il segnale con una banda utile tra 300 Hz e 3000 Hz. In seguito, un algoritmo di individuazione degli spike basato sulla Continuous Wavelet Transform è stato impiegato. Le caratteristiche degli spike così identificati sono state estratte applicando l'analisi delle componenti principali (PCA). Le prime tre componenti, che riescono a spiegare la maggior parte della varianza dei dati, sono state impiegate come input per l'algoritmo di clustering DBSCAN per determinare i neuroni.

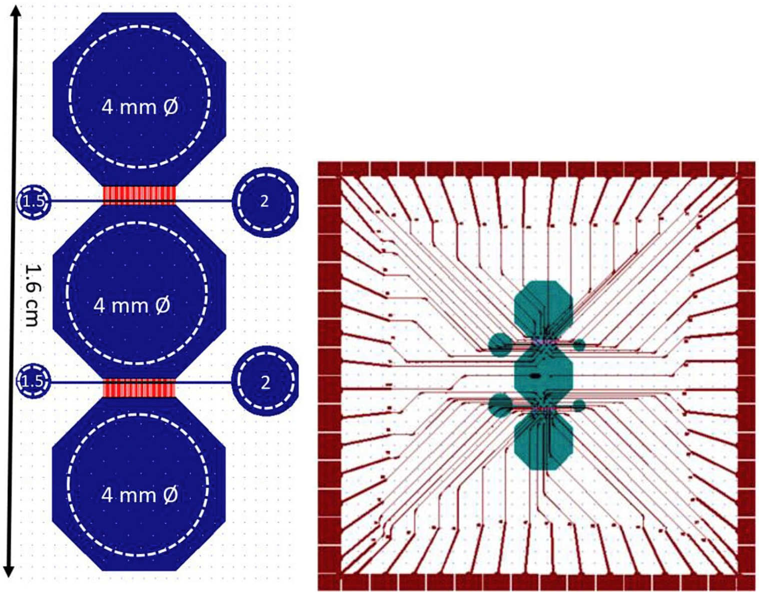

**Figure 1:** Rappresentazione grafica delle componenti del MEA impiegate. [Left] Design of microfluidic chip. [Sinistra] Design del chip microfluidico. [Destra] Design del MEA personalizzato con il chip in verde. [1]

In seguito alla procedura di spike sorting, l'intervallo di tempo tra due spike consecutivi (ISI) è stato calcolato. Infatti, l'attività extracellulare di una cellula neuronale può essere caratterizzata tramite la tecnica del Point Process modellizzando l'attività elettrica del neurone come una distribuzione statistica basata sui dati dell'ISI. In partcolare, distribuzioni stocastiche note in letteratura (Gaussiana Inversa, Gamma, Esponenziale, Log Normale) sono state valutate e i risultati esaminati con il test di Kolgomorov-Smirnov. I parametri della distribuzione sono quindi stati stimati usando l'algoritmo di sampling Hamiltonian Monte Carlo.

I parametri ottenuti sono stati impiegati per effettuare un'analisi multivariata per discrimiare i neuroni sani dalle cellule che presentato la mutazione del gene LRRK2 sia durante la fase di *Baseline* che dopo 24 ore dalla stimolazione. A tal proposito, quattro modelli di machine learning ((Decision Tree Classifier, Logistic Regression, Support Vector Machines, Random Forest) sono stati impiegati e le loro performance valutate.

## Risultati

L'algoritmo di spike sorting sviluppato è stato applicato per studiare i segnali bioelettrici delle reti neurali *in-vitro*. Inizialmente i dati sono stati filtrati applicando un filtro passabanda. Inoltre è stato necessario utilizzare un filtro notch per eliminare degli artifatti presenti nella banda di frequenza del segnale. L'algoritmo di identificazione degli spike basato sulla CWT è stato applicato per delineare gli istanti in cui accade uno spike, e, in seguito, la PCA è stata utilizzata per l'estrazione delle features, mentre la DBSCAN è stata impiegata per determinare i neuroni (figura 2).

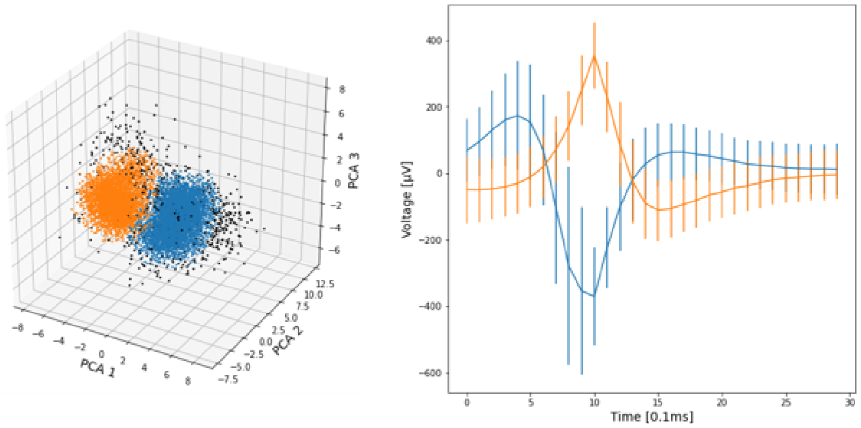

**Figure 2:** [Sinistra] Spike identificati mostrati nello spazio delle componenti principali. I cluster di spike delinenano due neuroni (cluster blu e arancione), e gli outliers (punti neri) sono identificati come rumore. [Destra] Forma d'onda media dei neurone identificati.

L'attività elettrica dei neuroni è stata quindi caratterizzata utilazzando l'ISI e la sua distribuzione modellizzata tramite il Point Process. Il test di Kolgomorov-Smirnov ha dimostrato che l'attività delle cellule identificate può essere descritta tramite una distribuzione Gaussiana Inversa (figura 3). Infatti, comparando la funzione di densità condizionata, il modello si mantiene all'interno dei limiti di significanza statistica (p-value$<$0.05). Quindi, i parametri della distribuzione per ogni neurone identificato sono stati stimati applicando l'algoritmo di sampling Hamiltonian Monte Carlo.

In ultima instanza, le differenze statistiche tra i neuroni mutati e sani sono state studiate. Una prima classificazione è stata effettuata per i dati

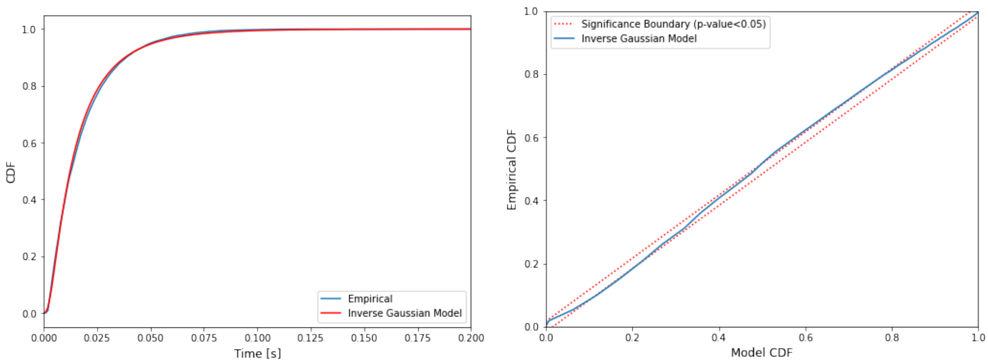

**Figure 3:** Distribuzione Gaussiana Inversa - grafici KS. [Sinistra] Confronto tra la CDF dei dati empirici (blu) e la CDF del modello (arancione). [Destra] Test di KS sul modello, in arancione i limiti di significatività (p-value<0.05).

ottenuti in riferimento alla fase di *Baseline*. A tal proposito, il dataset è stato diviso in un training e un test set con una proporzione di 75/25. I modelli di machine learning sono stati quindi impiegati per la classificazione e allenati con il training set. Essi sono risultati capaci di predirre in modo accurato i dati del test set con un picco di accuratezza di 96.15 ottenuto con il modello Random Forest e una rispettiva area sotto la curva ROC di 96.01 (figura 4).

In maniera analoga, una classificazione multi classe è stata effettuata per la fase dopo 24 ore dalla stimolazione. Per questa task si sono prese in considerazione quattro classi in cui i dati sono stati suddivisi: cellule sane controllo, cellule mutate controllo, cellule sane stimolato con KA, cellule mutate stimolate con KA. Il modello di machine learning per cui si è ottenuta la migliore performance è risultato il modello Random Forest con un'accuratezza di 74.8.

## Innovazioni

Un innovativo algoritmo di spike sorting è stato implementato per caratterizzare l'attività della reti neurali. L'identificazione degli spike è stata migliorata utilizzando un algoritmo basato sulla Continuous Wavalet Transform. Questo approccio permette l'indivduazione di spike aventi un picco di intensità limitato, basandosi sulla forma d'onda di un tipico potenziale

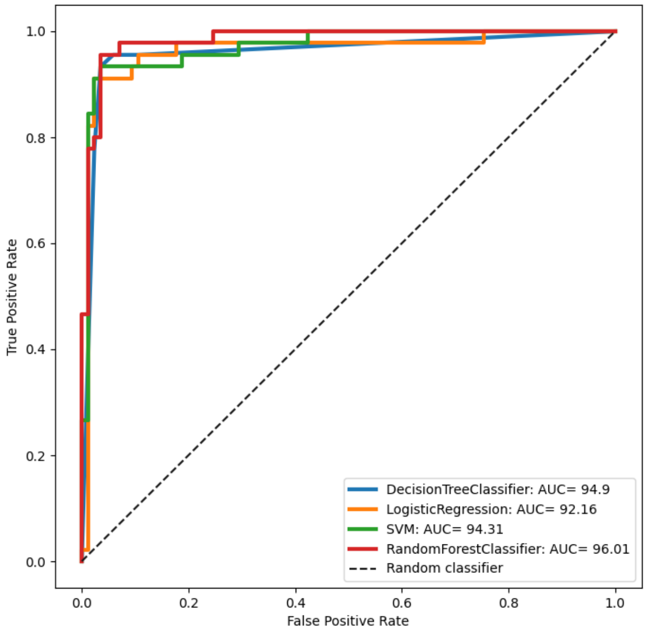

**Figure 4:** Curve ROC per la fase di*Baseline*.

d'azione registrato extracellularmente, ossia una madre wavelet biortogonale. Per cui, tale approccio non considera la statistica dell'intero segnale come nell'analisi sviluppata in precedenza, nella quale un metodo basato su una soglia fissa è stato impiegato, permettendo così una limitata sensibilità al rumore.

L'attività neuronale estratta è stata modellata come una distribuzione Gaussiana Inversa, la quale è risultata il modello migliore dopo l'analisi tramite il test di Kolgomorov-Smirnov. Tale distribuzione è considerata nello stato dell'arte come una buona stima dell'attività extracellulare di un neurone.[5]

Per il problema di classificazione, quattro modelli di Machine Learning sono stati utilizzati per valutare la differenza tra cellule sane e mutate. L'approccio di manipolazione del segnale implementato per questo lavoro migliora i risultati della classificazione in comparazione con l'analisi precedente.

Infatti, per la classificazione binaria durante la fase di baseline, l'accuratezza del modello di Random Forest raggiunge il valore di 96.15, mentre l'analisi precedente otteneva un valore massimo di 93.1. Tale miglioramento è visibile per tutti i modelli di classificazione presi in considerazione.

La classificazione multiclasse ha raggiunto performance migliori, con un'accuraztezza di 74.8 per il classificatore Random Forest, mentre il lavoro precedente otteneva un valore di 68.8 impiegando lo stesso modello.

### Discussione e Conclusione

Questo lavoro si pone l'obiettivo di fonrire una più accurata caratterizzazione statistica dell'attività di spiking registrata da reti neurali *in-vitro* aventi la mutazione G2019S sul gene LRRK2. Come descritto nel paragrafo precedente, questo studio migliora la caratterizzazione dell'attività elettrica delle reti neurali.

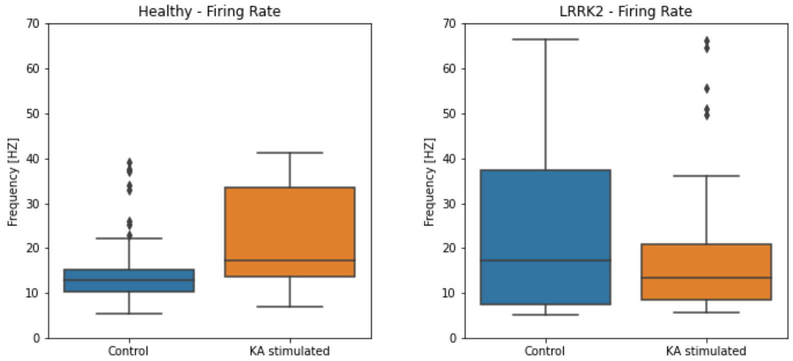

**Figure 5:** Frequenza di spike [HZ] durante la fase *dopo 24 ore* divisa secondo la popolazione di appartenza delle cellule.

Inoltre, l'analisi per la fase di *Baseline* ha mostrato una maggiore frequenza di sparo nelle cellule aventi la mutaziona sul gene LRRK2 rispetto alla popolazione sana di controllo. Tali risultati confermano ciò che è noto in letteratura.[6] Questa caratteristica è correlata ad una significativa differenza nel numero di mitocondri all'interno nelle cellule durante questa fase, che si ritiene la causa della distruzione del meccanismo di depressione a lungo termine delle sinapsi. [7][8]

In seguito alla stimulazione per mezzo dell'acido kainico, si osserva una diminuizione di tale valore. Questo comportamento può essere interpretato come l'inabilità per le cellule avanti la mutazione genetica di superare l'insulto della neurotossina.

**Parole chiave**: Multi-Electrode Array, mutazione LRRK2-G2019S, morbo di Parkinson, Spike Sorting, Point Process, classificazione con Machine Learning.

# Summary

### Introduction

In the last decades, *in-vitro* neural networks have been studied to assess the effects of genetic mutations related to neurological disorders on the correct bioelectric functioning of the cells.

The extracellular activity could be recorded by using Micro Electrode Arrays on which the cell cultures are placed. These devices comprise a series of small electrodes having dimensions of the order of hundreds of $\mu$m that are able to acquire the electric behaviour of a neuron with an high accuracy.

Furthermore, the development of new *in-vitro* techniques for cell growth after the discovery of Yamanaka's factor [2] have unlocked new opportunities for studying genetic mutations altering the status of human neural cells. In particular, human fibroblasts presenting a specific DNA mutation can be converted into induced-Pluripotent Stem Cells (iPSCs) from which the neural cells are derived. By integrating MEA technology and *in-vitro* neural cells, neurological disorder as Parkinson's Disease (PD) can be studied at a cellular scale.

Several studies[3][4] have shown how the G2019S mutation of the gene LRRK2 can be found in patients affected by PD. However, the role of this mutation on the pathology has not been fully delineated yet. Hence, the study of neural cells presenting this mutation can improve our understanding of this disease.

However, the acquired data needs to be processed in order to extract useful information about the behaviour of the neurons. To this end, spike sorting algorithms [9] can be applied to identify and characterize the neurons

from the bioelectric signals.

This work aims to develop a spike sorting pipeline employing a Continuous Wavelet Transform based spike detection algorithm to study the neural activity of LRRK2-mutated and healthy neurons. Furthermore, a Point Process framework is used to characterize the statistic of the neural behaviour of the cells. Finally, unsupervised classification will be performed in order to discern healthy and mutated cells employing the extract features.

This works stands as improvement of a first analysis performed on the dataset at Politecnico di Milano in collaboration with the Department of Neuromedicine and Movement Science at the Norwegian University of Science of Technology (NTNU) of Trondheim (Norway).

### Materials and Methods

*In-vitro* neural networks have been cultured starting from iPSCs derived from human fibroblasts by means of Yamanaka factors.[2] Hence, two groups are obtained such that six neural networks carry the G2019S mutation of the LRRK2 gene, while six comprise healthy neurons used as control. The neural networks have been cultured on a microfluidic chip composed by three interconnected chambers and placed on a custom MEA[1] (figure 6) having 60 electrodes. The spiking activity of the neural networks have been recorded on a *Baseline* phase and after 24 hours from a electro-chemical stimulation through Kainic acid.

The signals have then been analysed by means of a newly implemented spike sorting algorithm. First, a Butterworth bandpass digital filter of order 8 has been employed to filter the signals with a useful bandwidth between 300 Hz and 3000 Hz. Then, a spike detection algorithm based on the Continuous Wavelet Transform [10] have been employed. The features of the identified spikes have been extracted using the Principal Component Analysis. The three first components accounting for most of the variance of the data have been used as input of the DBSCAN clustering method in order to sort the spikes and identify the neurons.

After this step, the Inter-Spike Interval (ISI) sequence of each neuron

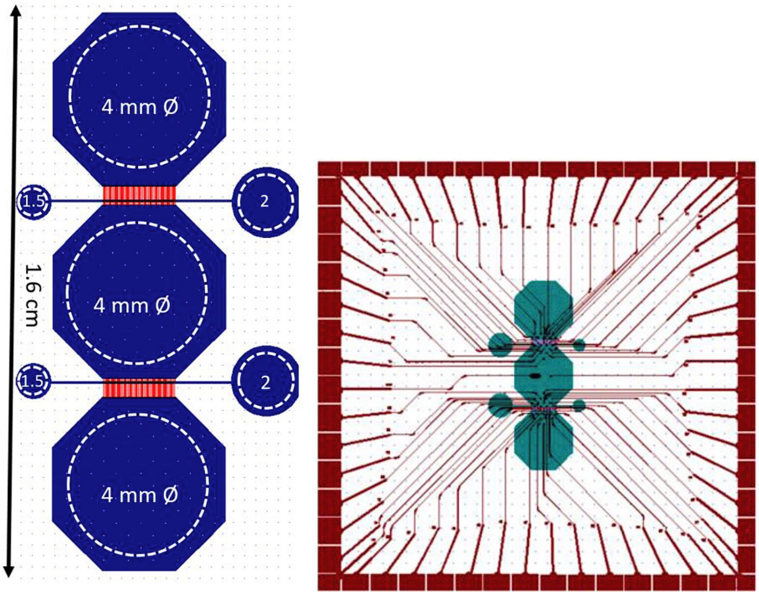

**Figure 6:** Graphical representation of the components of the MEA employed. [Left] Design of microfluidic chip. [Right] Design of MEA layout aligned with the chip design (green). [1]

is computed as the time between subsequent spikes. In fact, the extracellular activity of the neural cells can be characterized using a Point Process framework by modelling the electrical behaviour of the cells as a statistical distribution fitted on the ISI data. In particular, well known state-of-art stochastic distributions (Inverse Gaussian, Gamma, Exponential, Log Normal) have been assessed and the goodness-of-fit evaluated by means of the Kolgomorov-Smirnov test. The parameters of the distributions were estimated using the Hamiltonian Monte Carlo Sampling algorithm.

The obtained parameters have been used to perform a multivariate analysis in order to discriminate between healthy neurons and LRRK2-mutated cells both during the *Baseline* and the *After 24 hours* phase. For this purpose, four unsupervised machine learning models (Decision Tree Classifier, Logistic Regression, Support Vector Machines, Random Forest) were used and the performance assessed.

## Results

The developed spike sorting algorithm has been applied to study the acquired bioelectric signal of the *in-vitro* neural networks. First, the data were denoised out using the bandpass filter. Moreover, a notch filter has been necessary to address artifact components found in the useful frequency band. The CWT-based spike detection has identified the time instants of the spikes. Then, the spike features have been extracted through the PCA algorithm and the spike sorted using the DBSCAN clustering (figure 7).

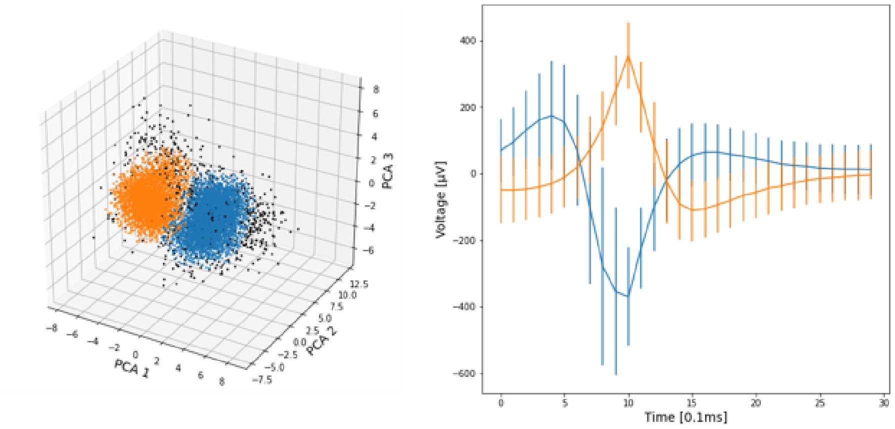

**Figure 7:** [Left] Detected spikes depicted according to their first three principal components. The neurons are clustered (blue and orange cluster), and outliers (black points) identified as noise. [Right] Mean waveform of each detected neuron.

The spiking activity of the neuron can be characterized by using the Inter-Spike Interval (ISI). Hence, the Point Process framework has been employed to model the distribution of the ISI sequences. The Kologomorov-Smirnov test has shown that the neural behaviour could be modeled by using an Inverse Gaussian distribution (figure 8).

In fact, comparing the empirical and the model Conditional Density Function, the model remains inside the statistically significant boundaries (p-value$<$0.05). Thus, the features characterizing this distribution have been estimated by using the Hamiltonian Monte Carlo Sampling algorithm for the identified neurons.

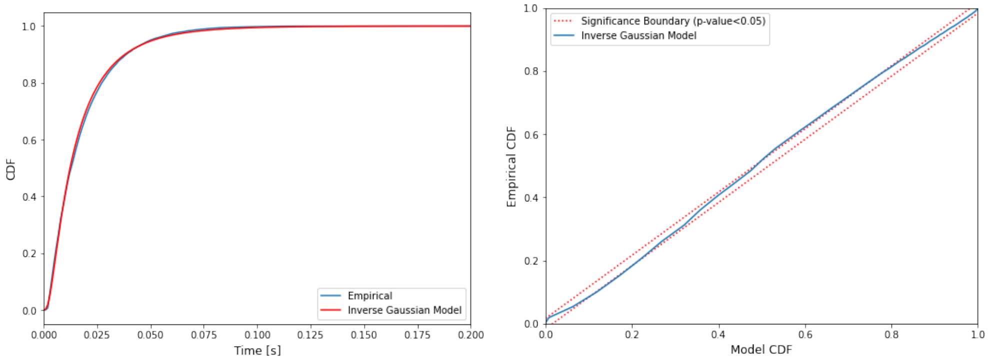

**Figure 8:** Inverse Gaussian model - KS plots. [Left] Comparison between empiral CDF (blue) and model CDF (orange). [Right] KS test on the model, in orange the significance boundaries (p-value<0.05).

Finally, the statistical differences between LRRK2-mutated and healthy identified neurons have been studied. First, a classification task was performed during the *Baseline* phase. To this end, the dataset has been divided between training and test set with a proportion 75/25.

Hence, the Machine Learning classification models have been trained and were efficiently able to classify the test set reaching an accuracy of 96.15 for the Random Forest classifier with an area under the ROC curve of 96.01 (figure 9).

In an analogous way, a multi-class classification has been performed for the *After 24 hours* phase. This task took into consideration four classes of samples: healthy control, LRRK2-mutated control, healthy KA-stimulated, and LRRK2-mutated KA-stimulated neural networks. The Machine Learning model with the best performance resulted the Random Forest classifier with an accuracy of 74.8.

## Innovations

A new spike sorting algorithm has been implemented to characterize the behaviour of the neural networks.

The spike identification step has been improved by using an algorithm based on the Continuous Wavelet Transform. This approach allows to identify spikes having low intensity peaks based on the shape of the waveform of a

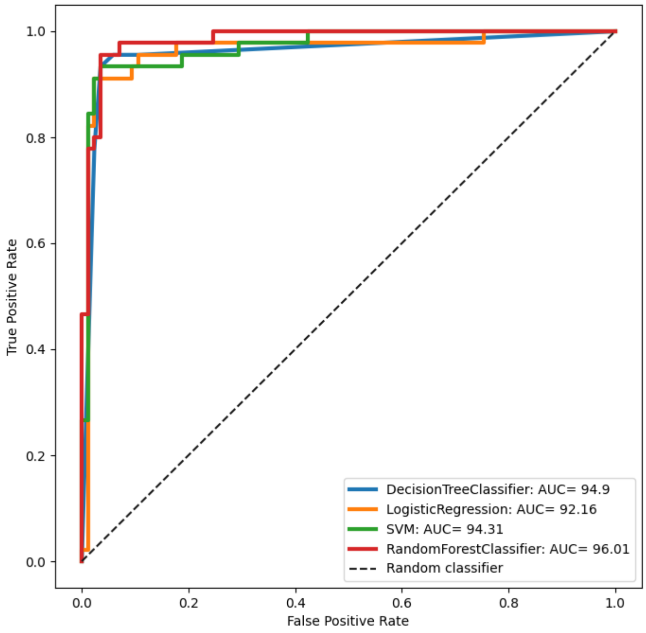

**Figure 9:** ROC curves for *Baseline* phase

typical extracellular record of an action potential, i.e. a biorthogonal mother wavelet. Therefore, the detection does not rely on the entire signal as in the previous work, in which an hard threshold based approach was employed, hence allowing a diminished sensibility to noise.

The extracted neural behaviour is successfully modeled using an Inverse Gaussian distribution, which resulted the best model among the stochastic distributions assessed after the Kolgomorov-Smirnov test. This distribution is considered by the state-of-art a good estimation of the extracellular activity of a neural cell.[5]

For the classification tasks, four Machine Learning models have been used to discriminate between healthy and mutated cells. The processing pipeline implemented in this work improves the results compared to the performance obtained in the previous analysis.

In fact, for the binary baseline classification, the accuracy of the Random

Forest classifier reaches 96.15, whereas the previous analysis obtained as best metric 93.1. This trend is visible for the other models as well.

The multiclass classification was improved with an accuracy of 74.8 for the Random Forest classifier, with the previous work obtaining 68.8 employing the same model.

## Discussion and Conclusion

This work aims to improve the statistical characterization of the spiking activity recorded from *in-vitro* neural networks carrying the G2019S mutation on the LRRK2 gene.

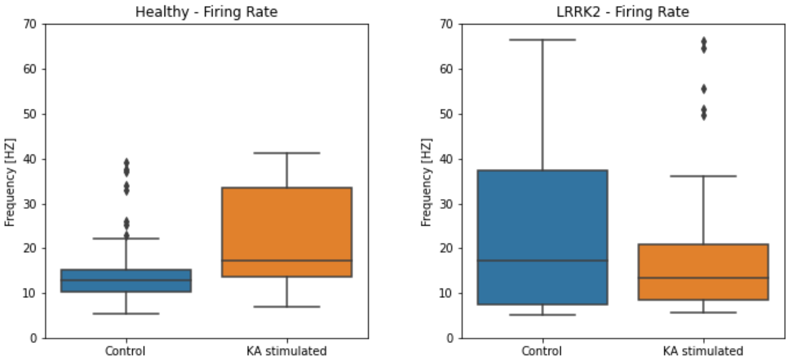

**Figure 10:** Firing rate [HZ] in the *After 24 hours* phase divided by groups

As outlined in the previous paragraph, several innovations were implemented, allowing a better characterization of the spiking activity of the neural networks.

Furthermore, the analysis at the *Baseline* phase has shown a greater mean firing rate in the LRRK2-mutated cells compared to the healthy control population, which corroborates the literature.[6] This activity is correlated to a significant difference in mitochondrial content within the cells during this stage, which is thought to disrupt the long-term depression mechanisms of the synapses.[7][8]

After the stimulation through Kainic Acid, a drop in this metric is observed in the mutated networks. This behaviour may be correlated to an inability

for the cells carrying the LRRK2 mutation to overcome the neurotoxic insult.

**Keywords**: Multi-Electrode Array, LRRK2-G2019S mutation, Parkinson's disease, Neuroscience, Spike Sorting, Point Process, Machine Learning classification.

# Chapter 1

# Introduction

In the last decades, the evolution of the technology in the neuroscience field has increased the opportunity to improve the knowledge of the nervous system. The advancements in the field have allowed to employ *in-vitro* neural networks for studying the pathology of neurological disorders by applying newly developed assessment techniques. In fact, even if *in-vitro* studies are not able to fully substitute the evaluation of the whole nervous system, they allow to focus on specific features that characterize the status of the cells. In this regard, new insights into neurological pathologies, such as Alzheimer's or Parkinson's diseases, have been discovered.

Parkinson's Disease (PD) is one of the most common neuro-degenerative disorder in the western world [11] and its impact is thought to be incrementing along with the ageing of the population. However, scientists are still far from a cure and the pharmacology can only slow down the chronic progression of the disease and reduce the symptoms.

Thus, the scientific community is studying the causes of this disorder at different levels. In particular, recent works have recognized that PD, such as several common age-related diseases, has a considerable genetic component. It has been proven that the gene Leucine-rich repeat kinase 2 (LRRK2) is a critical factor for understanding the mechanisms of this pathology[3], and of particular interest is the mutation G2019S.[4] For this reason, *in-vitro* neurons have been employed to assess the impact of this mutation and to study its effects on the cells, reproducing mutated neural networks and recording the electrical activity. Multi-Electrode Arrays (MEAs) have been shown to be a valuable method to study electrophysiological properties of the central nervous system neurons, allowing the recording of both *in-vivo* or *in-vitro* neural networks. Therefore, this technology has been widely employed in neuroscience. In fact, MEAs allow simultaneous and noninvasive cellular recordings from a large number of neural cells. [12] However, *in-vivo* applications can evaluate PD neural networks in animals, such as rats or primates, but it has not been employed in human subjects. Nonetheless, the study of *in-vitro* human neural cells affected by PD has proven to be an interesting researching field that may arise new knowledge on the mechanisms regulating neurodiseases.

This thesis will present an analysis of the electrical activity of *in-vitro* neural networks carrying the mutation G2019S for gene LRRK2 compared to a control population of healthy cells. The recording are provided by the Department of Neuromedicine and Movement Science of the Norwegian University of Science and Technology (NTNU) of Throdheim, Norway.[6] A previous analysis of the dataset [13] has shown statistical significant features of the neural activity of mutated-cells. This work aims to improve the results previously obtained by exploring more advanced methods in the spike activity assessment.

First, the recordings have been denoised using signal processing methods. Then, a spike sorting procedure had been employed to extract the spiking activity information from the signals. To this end, a Continuous Wavelet Transform based spike detection algorithm [10] has been used in order to obtain a more precise identification of the spikes. A non-parametric clustering algorithm, DBSCAN, have been employed to outline neurons by sorting spikes with similar characteristics.

Then, a Point Process analysis has modeled the neural behaviour extracting statistical features that characterize the spiking activity of the identified neural cells. The Hamiltonian Monte Carlo Sampling algorithm has been used for the parameters estimation by fitting a series of well known state-of-art statistical models with the empirical data. The Kolgomorov-Smirnov test has been employed to evaluate the goodness-of-fit.

Finally, a multi-variate analysis has been performed in order to study the obtained results for the different types of neural networks. Machine Learning models have been used to distinguish mutated and healthy cells at the *Baseline* phase and *After 24 hours* phase.

# Chapter 2

# State of the Art

Neuroscience is a multidisciplinary science that studies the nervous system in all its aspects; it combines technologies at micro and macro scale aiming to understand its fundamental properties. The nervous system is a highly complex component of the human body and it may be considered one of the latest anatomical and physiological system to be studied and understood by the scientific community. In fact, the last two centuries experienced an increasing interest in the field yielding to novel discoveries that changed our perception of the human neurophysiology.

However, this topic needs further studies and the scientific community is eagerly working to uncover the mysteries of the human mind. In the last decades several initiatives such as the US Brain Initiatives, the European Human Brain Project[14] or the Human Connectome Project [15] have pushed the research, but they have also shown that achieving a deeper understanding of exactly how the human neural system functions or the complete outlining of the etiology of neurodisorders still remain a profoundly demanding endeavor.

## 2.1 The study of *in-vitro* Neural Networks

### 2.1.1 Neural cells

A neural cell, or neuron, is the basic working unit of the nervous system; it is designed to transmit information through electric impulses to other nerve

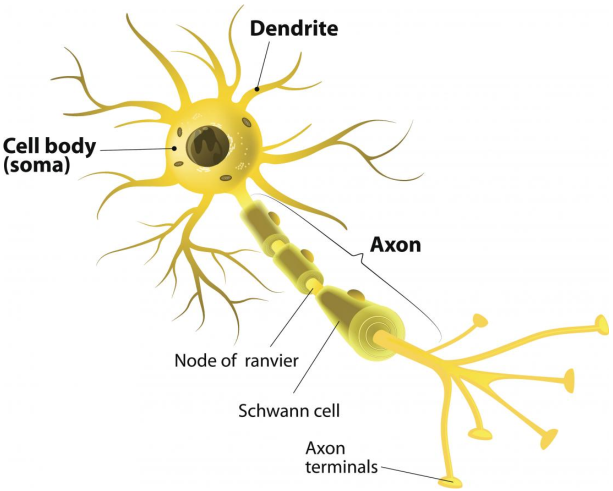

**Figure 2.1:** The neuron cell.

cells, or glands.

A neuron (figure 2.1) is typically composed by a cell body defined as soma, dendrites, and an axon, having different terminal ramification called synapses. The neuron is an excitable cell that generates an all-or-nothing impulse referred to as Action Potential (AP), which control the communication between two adjacent neural cells. This takes place in the synapse by means of an electrical stimulus, through a gap junction between the cells, or chemical stimulus, i.e. mediated by a neurotransmitter. In the soma, the incoming electric impulses are summed up temporally and spatially, and, if a certain voltage threshold is passed, the AP occurs.

In fact, AP is a fast variation in the voltage across the cell membrane elicited by the flow of specific ions. The cell membrane comprises a lipid bilayer that separates the intra-cellular and extra-cellular environments, that present a different concentration of charged ions, hence creating a voltage difference. The membrane may be thought as an electric capacity approximately equal to $1\mu F/cm^2$.

In resting conditions, the internal environment has an high concentration of potassium ions $K^+$, while a low concentration of sodium $N^+$. This status indeed creates a voltage potential across the cell membrane equal to -60mV and it is maintained as long as there is not a net ions flux. In fact, when a sufficient electric stimulus occurs the membrane potential changes and the so called sodium-potassium pump activates. This phenomenon consists in a positive feedback that yields to an incrementally opening of voltage-gated $N^+$ channels with a consequential rise of the membrane potential. The depolarization of the membrane is a rapid event that can last approximately 1 ms. Thus, the channels start to close while voltage-gated $K^+$ opens with the repolarization of the membrane. After the peak of the action potential, the neural cell presents a refractory period during which a new spike cannot be elicited.

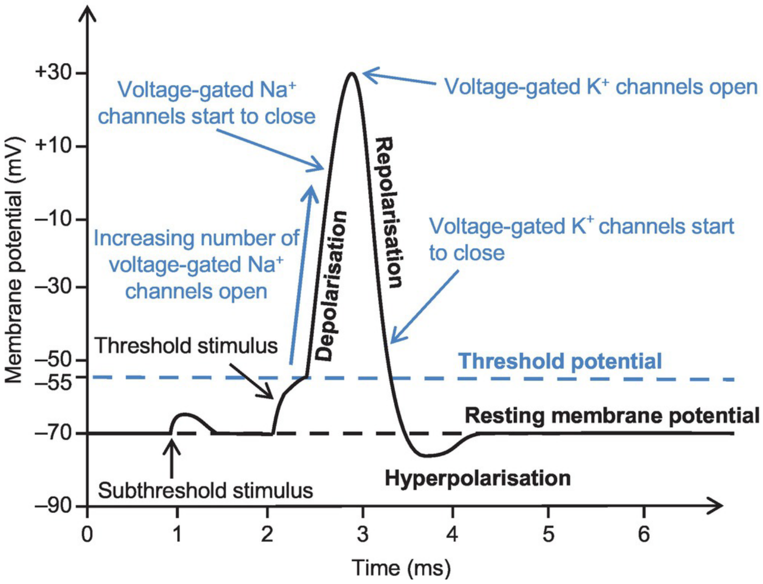

**Figure 2.2:** The Action Potential

An important aspect of the neural functioning is the synaptic communication. Synapses can be excitatory or inhibitory and the convey the afferent information to the neural cell respectively adding or decreasing the voltage potential. The weight of their contributions changes in time and this phenomenon is referred to as synaptic plasticity, which can strengthen or weaken the communication between two cells. This plays a key role in the early development of a biological neural network. This is an important phase studied in *in-vitro* cultures, because it is involved in specific tasks such as learning and memory.

### 2.1.2 Neural cells culture

*In-vitro* neural cells studies have thrived with the discovery of Yamanaka factors (c-Myc, Klf4, Oct3/4, Sox2).[2] This new technology have allowed to develop human neural network cultures starting from induced-pluripotent stem cells (iPSC). In fact, somatic cells, such as human fibroblasts, can be reprogrammed to become iPSCs through the viral overexpression of the Yamanaka transcriptors factor. This has yielded to efficiently recreate cultures of cells resembling the genetics of a specific subject, hence unlocking the opportunity to study the role of genetic patterns in a pathology. In this regard, iPSCs have been employed to study neurodegenerative disorders such as Alzheimer's or Parkinson's disease. A case of study that involves *in-vitro* cultures may in fact be the assessment of the bioelectric behaviour of human neural cells that have genetic mutations related to such disorders by means of multi electrode arrays.

### 2.1.3 Multi-Electrode Array

A Multi-Eletrode Array (MEA), also referred to as Micro Electrode Array, is an electronic device composed by a matrix of multiple micro-electrodes, which simultaneously can record the electric activity of several neurons. In fact, MEAs have been broadly employed in neuroscience to acquire extracellular spike activity of neural networks because of the non-invasiveness of this tool, which allows to extract both *in-vivo* or *in-vitro* records with no damage of the cell surface.

The first work that reported the employment of MEA was published in 1972 [16] and successfully provided a new framework for monitoring the bioelectric activity of *in-vitro* neural cells. From then, this technology has widely spread and further develop.

A typical MEA system is composed by an headstage and an interface board. The headstage is the core of the device and it has the purpose of housing the MEA plate, in which the probe and the cells are placed, recording the signals, and it may include a stimulation generator. Thus, it controls the acquiring process and its specifics such as sampling frequency, the amplification gain, or noise erasing, The interface board gathers the information from the headstage and transmits them to the main computer. It is equipped with a digital signal processor and it can be connected to various external instruments via multiple analog or digital inputs and outputs. The MEA probe is a plate composed usually by 60, 120, or 256 micro electrodes in planar titanium nitrile (TiN) spaced with a distance ranging from 100 to 500 $\mu$m. The probe is attached to the headstage.

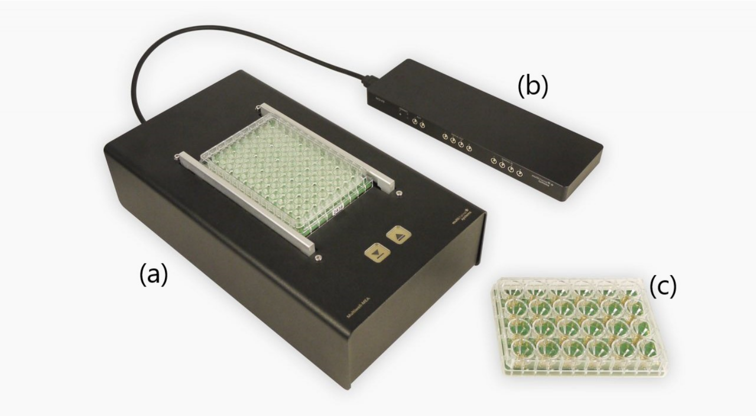

**Figure 2.3:** An example of MEA system by Multi Channel System. (a) Headstage, (b) interface board, (c) MEA probe chamber.

## 2.2 Parkinson's disease

Parkinson's disease is a chronic neurodegerative disorder of the central nervous system that affects the motor system inducing difficulties in the movements of the subject. Nowadays, the Parkinsons' disease is the second most common neurodegenerative disease (after Alzheimer's), and its incidence is comprised between 100-300/100.000 population, with an expected increment due to the general increasment of the life expectancy of the population.[11] The symptoms may vary between subjects but generally they are bradykinesia, rigidity, postural instability, resting tremors, and speech impediments.

Parkinson's desease is provoked by the loss of dopaminergic neurons in the substantia nigra, which leads to decreased levels of dopamine in the striatum and pathological modification in the circuitry of the downstream

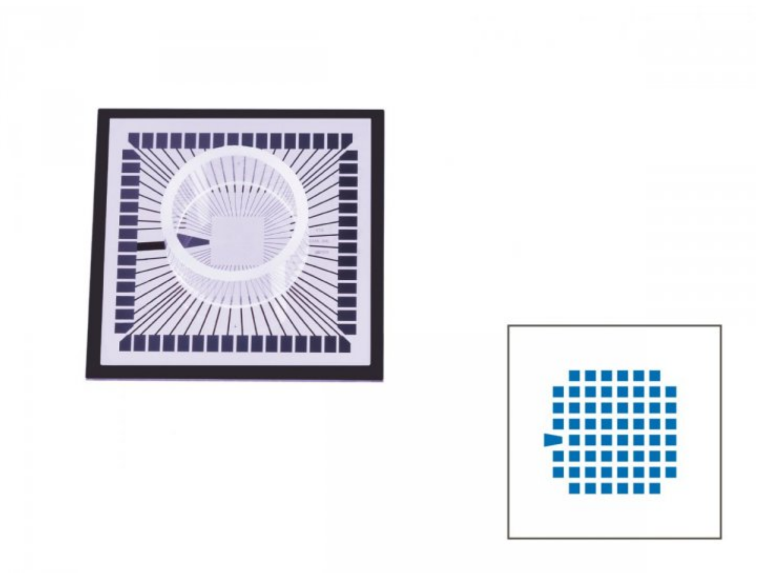

**Figure 2.4:** MEA probe with 60 electrodes by Multi Channel System.

basal ganglia.[17] Furthermore, this disorder is characterized by the presence of abnormal aggregation of protein inside the nerve cells, referred to as Lewy bodies, formed by aggregates of $\alpha$-synuclein (pre-synaptic protein).

### 2.2.1 LRRK2 gene mutation

Mutations of the LRRK2 gene has been found to cause PARK8-related Parkinson's disease, and it is considered the most common genetic cause of PD.[18] The LRRK2 gene is found in the twelfth chromosome of the human genome (cytogenetic location 12q12); it provides the information for encoding a protein called *dardarin* that is involved in the transfer of phosphate group from the energy molecule ATP to aminoacids.

Six of twenty mutations of the gene LRKK2 have been demonstrated to be pathogenic: the most common mutation is G2019S (6005G → A), which accounts for 5–6% of PD family-related and 1–2% of sporadic cases. [4] [3] It has been observed in *in-vitro* studies that the G2019S mutation consistently increases the kinase activity, impairing the protein stability and producing enlarged lysosomes with decreased degradative function.[19]

## 2.3 Modelling the neural activity

Since the first modelling definition proposed by Hodgkin and Huxley in the 1950s [20], the study of the neural activity has been strictly linked to the development of mathematical models aiming to characterize the neural electro-physiology and the interactions between cells. Nowadays, different techniques are available to extract information from the bioelectric activity of a neuron. However, several limitations are still present.

### 2.3.1 Neural activity recording

The acquirement process of the electric behaviour of neurons allows to extract cardinal information about the cell itself. It can focus on recording the intracellular or extracellular activity of the cell.

#### Intracellular recording

This approach aims to assess the voltage and the current in the internal environment of the neuron, and focuses in assessing the changes elicited by different stimuli. One of the most employed technique is the voltage-clamp, which measures the ion current flowing through the membrane cell, by maintaining the voltage constant. This allows to derive the values of the trans-membrane capacitance and provides information about the functioning of the channel gates. However, this is not enough to fully understand the behaviour of a neural cell interconnected in a network.

#### Extracellular recording

The electric activity of a neuron can be recorded externally, allowing to assess the behaviour of a cell (or a group of cells) both *in-vivo* and *in-vitro*. As introduced, Multi Electrode Arrays are widely employed to achieve this goal. However, it is hard to acquire the signal from a single neuron, since it would inevitably be corrupted by noise given by the acquirement device as well as the activity of the cells nearby.

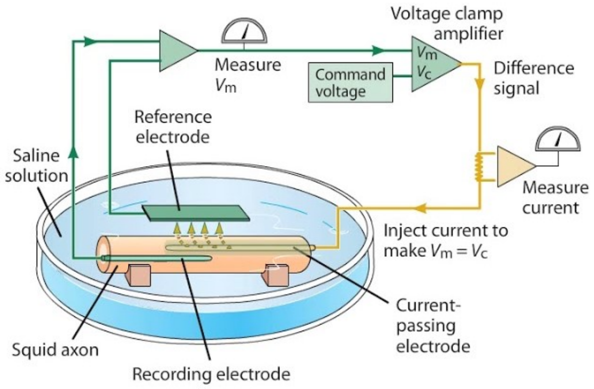

**Figure 2.5:** Voltage-clamp technique

### 2.3.2 Spike sorting

The problem of detecting transient signals, such as action potentials, in a noisy environment has been studied for decades. Measuring the activity of individual neurons accurately can be difficult due to large amounts of background noise and the difficulty in distinguishing the APs of one neuron from those of others in the local area. For experimental investigations it is ultimately important to accurately and robustly detect and localize the occurrence of individual spikes within the extracellular recording signal.[9][21]

The knowledge on the electrical activity of networks of neurons has improved introducing new technologies and algorithms. To this end, spike sorting is a crucial step to gain information from extracellular neural activity recording. Nonetheless, a common unique reference technique has not been established, leading to the development of different approaches. However, every spike sorting pipeline comprises of different steps that can be summarized as:

1. Filtering
2. Detection
3. Feature extraction
4. Clustering

Extracellularly recorded spike trains are inevitably corrupted by noise. The noise sources are quite varied: the recording hardware, the ambient recording environment, the superimposed activity of multiple neurons, and the spatially averaged activity of distant cells also known as the local field potential. Perhaps most importantly, the activity of distant neurons may appear as noise which is highly correlated with the useful signal. The following paragraphs will introduce an overview for each step of the spike sorting procedure, showing the state of the art for each of them.

### Filtering

The extracellular electric signal recorded from neurons is highly noisy, thus it is necessary a filtering step before starting the actual spike sorting procedure. In fact, the raw data is given by the sum between the useful signal, i.e. the spiking activity, and the background noise, given by the Local Field Potentials (LFPs). Thus, a digital band-pass filter is applied to clean the signal, usually between 300 Hz an 3 kHz. Moreover, the power spectrum of the signal is usually assessed to evaluate the presence of noise which may be given by the hardware acquisition of the signal.

### Spike detection

Spike detection algorithms aims to identify the time instant in which a spike occurs. Different approaches may be used. A widely used technique for spike detection is amplitude thresholding, where the threshold value can be set automatically, for instance as a multiple of the estimated noise standard deviation, or manually. Although this detection method is simple, its performance decreases under low signal-to-noise ratio (SNR) conditions.

Other detection methods include power detection, matched filtering, principal components, Haar transformation, and the discrete wavelet packet transform. In the power detection method, also known as energy detection, the instantaneous power of the signal, calculated using a sliding window approach, is compared against a threshold derived from the mean and the standard deviation of the noise power.

In particular, a focus on hard threshold based and Continuous Wavelet Transform based algorithm is here provided.

- Hard threshold techniques: generally, the most prominent feature of the spike shape is its amplitude. One of the simplest ways to measure the activity of a neuron is by defining a voltage threshold above which the spike is detected. This approach requires minimal hardware and software, however it is not always possible to achieve acceptable isolation and a noisy environment can affect the results. Thus, the threshold level determines the trade-off between missed spikes (false negatives) and the number of background events that cross threshold considered as spikes (false positives).

- Detection of action potentials with the continuous wavelet transform: Continuous wavelet transform (CWT) can be used to study the signal in both time and frequency domain.[10] The duration and the waveform of a typical spike event is well known in literature. This *a-priori* information allows to zoom in on the scales of interest improving the spike detection. Furthermore, the wavelets of compact support are preferred, because they resemble the nature of APs. In particular, wavelets from biorthogonal family have bi-phasic shapes that are reminiscent of those of neural spikes. This allows a single AP to be represented with a few wavelet basis functions—sparse representation.

### Features extraction

Once the instants of the spikes are detected, features needs to be extracted to identify the neurons. The most common algorithm employed for this step consists in the Principal Component Analysis (PCA). Moreover, wavelet functions may be used by considering the wavelet coefficients as the features characterizing the spike.

### Clustering

The extracted features are then clusterized in order to define the neurons. To this end, unsupervised machine learning algorithms are employed to sort the spikes. In this way, the spikes are assigned to a cluster based on their features and the neuron defined. Widely diffused clustering methods are non-parametric algorithms based on nearest neighbor interactions.

## 2.4 Analysis of spiking activity from MEA

MEA systems have been widely employed in studying activity of *in-vitro* neural networks. However, in order to efficiently extract information about the neural behaviour using this technology, an efficient process pipeline needs to be defined. In addition to the spike sorting procedure, statistical methods can be applied to characterize the behaviour of the cells starting from the recorded signals.

Most information about the nature of the bioelectric activity are thus extracted from the spike time-series, such us statistical characteristic as the firing rate and the distribution of spikes in terms of time interval between two consecutive APs (Inter-Spike Interval), mean, median, standard deviation.

Furthermore, the burst activity has been considered as an important aspect in the inter communication between cells inside a network, especially during the connections development stage.[22] Bursts are defined as spikes occurring at higher frequency and could be detected by studying the Inter-Spike Interval.[23]

### 2.4.1 Previous works

Previous analyses has been performed on the dataset studied in this work.

Valderhaug et al., 2020,[6] have observed alterations regarding both structure and functioning of the human neural networks carrying the G2019S mutation. They have provided first evidence of increased neuritic density in mutated population, along with increased baseline spiking activity compared to healthy control neurons. Finally, they have observed different responses to overexcitation through Kainic Acid. In particular, LRRK2-mutated neural networks have shown a great drop in average mean firing rate after 24 hours from the neurotoxic insult.

| ML model | Accuracy | Recall | Precision | ROC AUC |
|---|---|---|---|---|
| DT | 85.34(87.93) | 90.20(82.35) | 79.31(89.36) | 85.87(87.33) |
| LR | 80.17(81.03) | 66.67(66.67) | 85.00(87.18) | 78.72(79.49) |
| SVM | 87.93(87.10) | 80.39(78.40) | 91.11(90.90) | 87.12(86.10) |
| RF | 93.10(88.80) | 86.27(84.30) | 97.78(89.60) | 92.37(88.30) |

**Table 2.1:** Baseline classification.

| ML model | Accuracy | Recall | Precision |
|---|---|---|---|
| DT | 54.65(46.51) | 54.65(46.31) | 53.87(25.77) |
| LR | 55.81(56.98) | 54.64(55.55) | 55.49(56.33) |
| SVM | 65.12(61.63) | 64.22(59.59) | 64.85(60.97) |
| RF | 68.60(58.14) | 67.62(56.87) | 68.15(57.88) |

**Table 2.2:** After 24 hours classification.

Furthermore, a thesis work developed at Politecnico di Milano in collaboration with the Department of Neuromedicine and Movement Science of the Norwegian University of Science and Technology (NTNU) of Throdheim, Norway,[13] has analysed the signals acquired from the neural networks. The main features of the approach used are here described. In the spike sorting, a spike detection HT-based approach with a threshold equal to four times the Median Absolute Deviation (MAD) (equation 2.1) has been employed.

$$MAD = \pm 4 \, median(|x_i - \bar{x}|) \tag{2.1}$$

A Dirichlet Mixture model has been employed for the statistical characterization of the identified neurons. In particular, the distribution is composed by an Inverse Gaussian distribution and two normal distribution having the mean respectively equal to 50 ms and 150 ms. Nine parameters have hence been estimated.

Finally, a binary classification on the baseline phase and a multi-class classification after 24 hours have been performed with the results shown in Table 2.1 and Table 2.2.[13]

# Chapter 3

# Material and Methods

## 3.1 Neural networks

Two groups of structured cortical neural networks have been cultured starting from human inducted-Pluripotent Stem Cells (iPSCs)-derived H9N Neural Stem Cells (NSCs). [6] In particular, the two groups are composed respectively by:

- 6 healthy neural networks, used as control (ax0019),
- 6 mutated networks carrying the LRRK2 G2019S (GGC>ACG) mutation (ax0310).

The iPSCs has been derived from dermal fibroblasts through Episomal iPSC (oriP/ EBNA1) Reprogramming Vectors. The full development protocol is avalaible in [6]. A resting-state recording of the electrical activity has been performed for 400 seconds for each group. Those data are referred to as *Baseline*.

### Kainic Acid stimulation

Fifteen days post seeding, three networks out of each group have been electro-chemically stimulated with Kainic Acid (KA) and the activity recorded. The acid was applied in the top cell chamber for 30 minutes. Kainic acid is used to define a subgroup of receptors for the excitatory neurotransmitter glutamate. These receptors inhibit fast synaptic transmission in the central nervous system and gate an ion channel regulating the influx of sodium ions. Kainic acid is a highly potent receptor agonist, and as such in this system, it is a neurotoxin and excitant.[24] Thus, the chambers of the networks stimulated through KA were washed out using Dulbecco's phospahte buffered saline (PBS) to remove the acid and stabilize pH. The control neural networks were also washed with PBS as a control condition. After this procedure, additional recordings have been extracted and defined as *After Stimulation* phase.

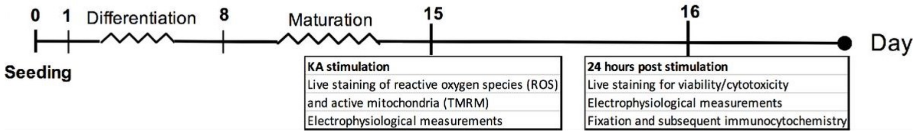
0
1
Differentiation
8
Maturation
15
16
Day
Seeding
KA stimulation
Live staining of reactive oxygen species (ROS)
and active mitochondria (TMRM)
Electrophysiological measurements
24 hours post stimulation
Live staining for viability/cytotoxicity
Electrophysiological measurements
Fixation and subsequent immunocytochemistry

**Figure 3.1:** Time schematic of culture and recording phases.

## 3.2 Multi-Electrode Array - MEA

The neural cells previously presented have been seeded on 3-nodal polydimenthylsiloxane (PDMS) microfluidic chips bonded to multi-electrode arrays, having a custom designed as shown in figure 3.1. The neural cells are placed in the octagonal node having a diameter of 4mm, which are interconnected by 104 microtunnels containing axons and dendrites developed from the soma of the cells placed in the closed chambers (nodes).

The microfluidic chips are placed above 59 recording microelectrodes plus a reference electrode. The activity of the neural networks has been recorded using the MEA2100 workstation at a sampling rate of 10 kHz, for an interval of time between 5 and 10 minutes. [1][6]

## 3.3 Spike sorting

The acquired electrical signals are then processed in order to detect the extracellular spike events generated by the neurons. The spike sorting

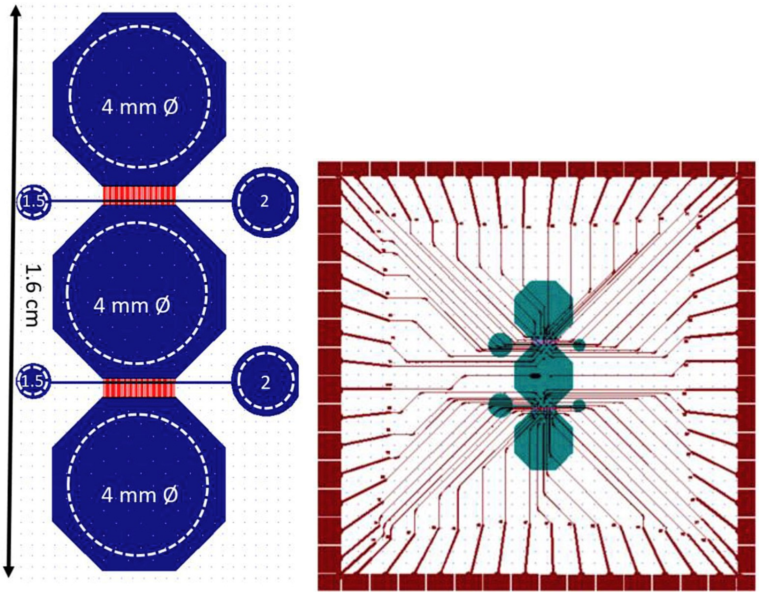

**Figure 3.2:** Graphical representation of the components of the MEA employed. [Left] Design of microfluidic chip. [Right] Design of MEA layout aligned with the chip design (green). [1]

technique is thus employed to extract such information. This paragraph will described in detail the methods applied to achieve this goal.

### 3.3.1 Filtering

Digital signal filtering aims to separate noise, especially given by Local Fields Potentials (LFPs), from the the electrical activity of a single neural cell. It is known in literature [9] that LFPs are recorded with a frequency range of [0.5-100] Hz, while the useful signal, i.e. the spiking activity, is typically found between 300 Hz and 3000 Hz.

Thus, a band-pass Butterworth filter can be used to erase the noise components. This type of filter corresponds to an infinite impulse response filter (IIR) that is maximally flat corresponding to the bandpass frequency range. A Butterwoth filter of order 8 is employed: a forward-backward filtering has been performed to avoid phase distortions given by non-linear phase responses.

Moreover, in order to assess the frequency components of the raw signal, its Power Spectral Density (PSD) has been computed such that

$$PSD(f) = \frac{1}{N} \left| \sum_{n=1}^{N} x_n(t = n\Delta t) e^{-i2\pi f n\Delta t} \Delta t \right|^2 \tag{3.1}$$

with N number of samples, $\Delta t$ sampling step, and $x$ the discrete signal. This procedure allows to verify the presence of artifacts in the signal that, eventually, will be denoised using a notch filter.

### 3.3.2 Correlated channels

One of the purposes of *in-vitro* MEA consists in providing a detailed spatial location of single neurons of the neural net. If the recording channels are sufficiently well separated, then there is no or little overlap between their signals, and spike sorting can be performed. However, often electrodes are packed in a limited region, with a distance between them that could compromise the isolation of each channel. In fact, signals recorded by close electrodes may result high correlated, with electrodes recording simultaneously the same spike activity.

Therefore, it is necessary to assess the correlation between the time-series of the channels. To this end, the Pearson correlation between the channels is pairwise computed as

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} \tag{3.2}$$

where $\sigma_{XY}$ represents the covariance between the two time-series $X$ and $Y$, $\sigma_X$ and $\sigma_Y$ are the variances of the two signals, and N is the number of samples for each channel. Channels depicting a correlation of $\rho_{XY} > 0.85$ were considered as highly correlated, and are replaced by the channel having the highest absolute maximum value.

### 3.3.3 Spike detection

The spike detection step has been implemented using a Continuous Wavelet Transform (CWT)-based approach.[10] The obtained results will be compared to an hard threshold spike detection that used the Median Absolute Deviation (MAD), which was used in a previous work.[13] The algorithm can be divided in five sections:

1. Multi-scale decomposition of the signal using a wavelet basis
2. Noise separation at each scale from useful signal
3. Bayesian hypothesis testing performed at each scale to evaluate the presence of a spike event
4. Combination of decisions at different scales
5. Estimation of event time instant

First, the signal is projected onto a set of basis functions defined by a set of expansion coefficients in the time-frequency domain. In particular, there exist a set of wavelet basis functions well suited for spike detection, such as Haar, Daubechies, or Biorthogonal wavelets. A wavelet $\psi$ is a function having zero average and finite energy.

$$\begin{gathered}\int_{\mathbf{R}} \psi(t)dt = 0 \\ \psi \in L^2(\mathbf{R})\end{gathered} \tag{3.3}$$

The function is normalized, i.e. $\|\psi\|_2 = 1$ and centered, and it is referred to as *mother wavelet.* From it, it is possible to obtain a family of time-scale waveforms applying transitional or scalar transformations such that:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \text{ with } a, b \in \mathbf{R}, \tag{3.4}$$

where $a > 0$ is the scale factor, whereas b is the translation factor. All the functions $\psi_{a,b}$ share the properties of the mother wavelet. The wavelet transform $X$ of a function $x \in L^2(\mathbf{R})$ is its projection onto the wavelet basis $\psi$ such that

$$X(a,b) = \int_{\mathbf{R}} x(t)\psi(t)dt. \tag{3.5}$$

Given scale $a_0$ and translation $b_0$, the wavelet transform of the function $x(t)$, i.e. $X(a,b)$, represents a similarity index to the wavelet $\psi_{a_0,b_0}$, and it is called *wavelet coefficient*. The biggest this value, the strongest is the similarity between the signal and the wavelet. For wavelet of compact support the domain of integration $\mathbf{R}$ is limited to a specific interval of time. Consequently, the wavelet coefficient will depend only on the signal within the time window considered.

Moreover, if the scale of the wavelet function can be considered small, the wavelet transform is able to capture transient signal phenomena. Therefore, the above mentioned characteristics make this approach very suitable for spike detection. To this end, the wavelet transform of the extracellular electric signal of the neurons determines the index of its resemblance with a typical spike waveform, and the time window, known *a-priori* from the literature, is defined between 0.5 ms and 2 ms.

### Wavelets

The choice of the mother wavelet is important to detect the spike event despite the background noise. In fact, the wavelet functions have to present similarity with the typical spike shape. In figure 3.3, the wavelet proposed by Nenadic [10] are depicted.

The orthogonal wavelets can be constructed from a single basis set, while the biorthogonal wavelets are constructed from different ones. The wavelets comprising the biorthognal family are designated in use as *biorNR.ND*, where "bior" stands for biorthogonal, "NR" and "ND" stand for the effective number of the reconstruction filters and effective number of the decomposition filters respectively. Because the wavelets project from different sources, filters which are on the other hand weightings of the scaling and wavelet functions must as well be different in each case so that biorthogonal wavelets differ from the orthogonal wavelets that use the same filters. [25]

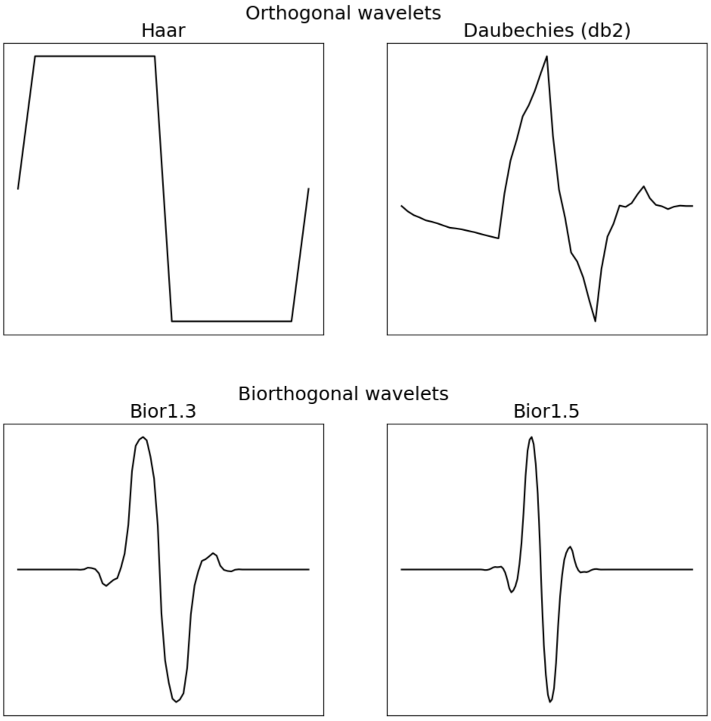

**Figure 3.3:** Examples of wavelet used for spike sorting. [Top] Orthogonal wavelets, Haar and Deubechies. [Bottom] Biorthogonal wavelets.

If compared to spikes, it is possible to notice how biorthogonal wavelets better resemble the actual waveform, as they present a bi-phasic shape typical of APs. Thus, it is expected that wavelets of this family will provide a sparser representation of neural signal than orthogonal ones, such as *db2*, allowing to describe the signal with few parameters.

### Choice of scale

The set of basic function translations $B$ is determined by the sampling rate of the signal $f_s[kHz]$ and its duration $T[s]$, such that $b \in B$, where $B = \{0, 1, ..., k, N-1\}$, and $N = Tf_s + 1$ corresponds to the number of samples composing the signal. Thus, the set of translations is formed by the discrete time vector. On the other hand, the set of scales $A$ is determined starting from the $a-priori$ definition of a spike duration interval. In literature it can be found that APs last between 0.5 ms and 2 ms. Given this fixed time window, the set of scales $A = \{a_0, a_1, ..., a_j, ..., a_J\}$ is defined as a vector of uniformly sampled values, starting from $a_0 = 0.5$ms and $a_J = 2$ms.

### Wavelet coefficients

By applying the CWT with the set of scales and translations previously defined, a multi-scale representation of the signal in terms of its wavelet coefficients is obtained. It is possible to define the discrete signal $x$ as a sum between the useful signal $s$ and a noise component $w$, such that

$$x[n] = s[n] + w[n] \quad n \in B. \tag{3.6}$$

Thus, given the signal representation using the wavelet transform, the coefficients $X(a_j, b_k)$ will be zero-mean random fluctuations in case of noise, while random variables having means different to zero when the useful signal $s$ is present. To this end, a threshold $\Theta$ on the value of the wavelet coefficients is defined as

$$\Theta_i = \sigma_i \sqrt{2 \log_e N}, \tag{3.7}$$

with $N$ number of samples of the windowed signal, and $\sigma_i^2 =$Var$\{X(a_j, b_k)\}$. Since the noise may be not white, the threshold becomes dependant on the basis function, such that each wavelet coefficient has a specific threshold. The variance $\sigma_i$ is estimated from the observed coefficient as the median of the absolute deviation of the coefficients for a specific scale j

$$\hat{\sigma_j} = median\{|X(j,0) - \bar{X}_j|, ..., |X(j, N-1) - \bar{X}_j|\}/0.6745, \tag{3.8}$$

with $\bar{X}_j$ sample mean of $X_j$.

### Detection algorithm at a single scale

The problem of spike detection is formulates as a binary hypothesis test, where the null hypothesis $\mathcal{H}_0$ states that only the noise $w$ is present, whereas for $\mathcal{H}_1$ both signal $s$ ad noise are present.

$$\begin{aligned} &\mathcal{H}_0 : x[n] = w[n] \quad n \in B \\ &\mathcal{H}_1 : x[n] = s[n] + w[n] \quad n \in B \end{aligned} \tag{3.9}$$

Thus, the two hypothesis are tested using the rules following

$$\begin{aligned} &\mathcal{H}_0 \;\text{ if }\; |X(a_J, b_k)| > \Theta_j \\ &\mathcal{H}_1 \;\text{ if }\; |X(a_J, b_k)| < \Theta_j \\ &\Theta_i \triangleq \frac{\hat{\mu_j}}{2} + \frac{\hat{\sigma_j^2}}{\hat{\mu_j}} \log_e \gamma_j \;\; \forall k \in B \end{aligned} \tag{3.10}$$

where $\hat{\sigma_j}$ as in equation 3.6, $\hat{\mu_j}$ is the sample mean of the absolute value of the wavelet coefficients at scale $a_j$ under the hypothesis $\mathcal{H}_1$, and $\gamma_j$ is a cost coefficient.

In order to define the cost parameter, it must be specified the value of a parameter $L \in [-0.2, 0.2]$, which define the trade-off between omissions and false detected spikes. $L > 0$ increases the number of false positive, on the other hand $L < 0$ increases the number of spikes not detected.

$$\log_e \gamma_j \triangleq LL_M + \log_e \frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1)} \tag{3.11}$$

with $L_M = 36.7368$, and the ratio between the two probabilities of each hypothesis $P(\mathcal{H}_0)/P(\mathcal{H}_1)$ is determined as the ratio between the number of the wavelet coefficients below the threshold for the scale *j-th* for each translation $b \in B$ (no spike detected), and the number of spike detected, i.e $X(j, k) < \Theta_i$, $\forall k \in B$.

### Combining the decisions of each individual scale

Because they are highly localized in time, the samples corresponding to neural transients occupy contiguous subsets of the discrete time vector. This property of transients is referred to as a *temporal contiguity.* Temporal contiguity translates into the contiguity of the coefficients in the wavelet domain, i.e. the wavelet coefficients corresponding to the same transient tend to be neighbors in both time and scale. The scale contiguity can also be viewed in the present context as a cross-correlation (redundancy) of the wavelet coefficients at different scales.

#### Estimation of spike time instant

In a noise-free environment, the wavelet basis function that provides the maximum correlation with the transient to be detected, corresponds to a wavelet coefficient of maximum magnitude. The time associated with the translation index of the basis function with maximal coefficient can be taken as a good approximation to the occurrence time of the event.

However, in a noisy environment, there is a jitter associated with the location of this maximal coefficient, which can be reduced by averaging the locations of the maxima across different scales. Finally, the highest absolute value of the signal in a window of 0.5 ms, chosen as tolerance window, centered on the identified spike instant is picked as the maximum of the spike wave.

### 3.3.4 Features extraction

For each detected spike a window of 30 samples (equal to 3 ms) is extracted to characterize the spike waveform, where the spike maximum (or minimum) is placed at the 10th sample of the window, e.g 1 ms. Thus, the Principal Component Analysis (PCA) is employed to extract the characteristic features of each waveform. PCA is an orthogonal linear transformation of the data space, able to transpose the data into a reduced space described by orthogonal basis, i.e. the principal components, along which the highest amount of variance of the datataset is explained.

This procedure allows to reduce the size of the dataset maintaining the information carried by the data by creating a lower-dimensional projection of the dataset. Before applying the PCA process, the data need to be standardized first. To this end, the z-score is computed as

$$x_{adj} = \frac{x - \mu}{\sigma} \tag{3.12}$$

where $\mu$ represents the mean of the data and $\sigma$ its standard deviation. Having a set of standardized data $X$, the first principal component is computed by solving the optimization problem of equation 3.13. The vector $w_1$ corresponds to the weight vector that characterizes the component along which the data are transposed maximizing the variance explained. Mathematically, this means

$$\underset{w_1}{max}\left\{\frac{w_1'Vw_1}{w_1'w_1}=1\right\} \tag{3.13}$$

where $V$ is the covariance matrix of $X$. The following principal component are computed by solving the same optimization problem, with the addition of the orthogonality constriction with the previous identified component, such that

$$\underset{w_2}{max}\left\{\frac{w_2'Vw_2}{w_2'w_2}=1\right\}, \quad w_2'w_1=0. \tag{3.14}$$

For the purpose of this work, the first 3 principal components are considered.

### 3.3.5 Clustering

The detected spikes were grouped in clusters using the Density-Based Spatial Clustering of Application with Noise (DBSCAN) method to identify the different neurons recorded by each channel. DBSCAN is a non-parametric clustering algorithm which is effectively able to identify dense clusters within a dataset, marking as outliers those points lying in low-density regions of the space.

This algorithm uses two input parameters: $\varepsilon$ corresponds to the maximum distance that defines two close points as belonging to the same clusters, whereas $MinPts$ refers to the minimum number of samples that must be enclosed into a cluster in order to define it as one. Considering a set of point, we define

- *Core point* the point $p$ if there are at least *MinPts* - 1 having a distance lower than $\varepsilon$ from $p$.
- *Directly reachable point* a point $q$ such that $dist(p, q) < \varepsilon$.

- *Reachable point* a point $q$ such that, given a path $p, p_1, ..., q$, each $p_{i+1}$ is directly reachable from $p_i$.
- *Outliers or noise points* every sample not reachable.
- Two points $p$ and $q$ are *density-connected* if there is a point $k$ such that both $p$ and $q$ are reachable from $k$.

Given that, the algorithm works as follow:

1. An arbitrary core point $p$ is chosen.
2. The *density-reachable* samples from p are evaluated.
3. If the number of *density-reachable* points is greater than $MinPts$, a cluster $C_1$ is identified.
4. The procedure is repeated from a different arbitrary point q $\notin C_1$, identifying the cluster $C_2$.
5. If the inter-cluster distance, defined as the minimum distance the points assigned to the different clusters, is less than $\varepsilon$,

   $$min\{dist(p,q)|p \in C_1, q \in C_2\} < \varepsilon, \tag{3.15}$$

   $C_1$ and $C_2$ are merged in the same cluster.

Thus, for each channel the neurons were identified by grouping spikes having similar features. Moreover, this approach removes false detection of spiking events by classifying spikes with no similarity with other events as noise.

## 3.4 Modelling

Once the spike sorting procedure has been completed, it was necessary to characterize the spiking activity of the single neuron, in order to outline differences between the control population and the mutated cells. Thus, the Inter-Spike Interval (ISI) of each neuron has been used to characterize the identified cell. Them, two different approaches were pursued to model the statistical features of each ISI.

### 3.4.1 Point Process Framework

A point process is defined as a binary stochastic process occurring in continuous time or space. Therefore, the spiking activity of neurons can be modeled using a temporal point-process. In fact, the binary representation of information about a time series is efficiently able to characterize the neural activity by simply using the spiking times of the events for each neuron. A point process can be represented either by the timing of the spikes, i.e. by the interval between the events, or as a set of binary values (1 if a spike has occurred, 0 otherwise). Thus, given an ordinate set of spike train $\mathbf{T}$, where $0 \leq T_1 \leq T_2 \leq ... \leq T_k$, a process $\mathbf{X}$ can be represented either:

1. **Counting measure**: the number of events $N_x$ over the interval of time [0, t], given by
$$N_x(0,k) = \sum_{k=1}^{\infty} I(T_k \leq t) \tag{3.16}$$
where $I(s) = 1$ if $s$ is true, 0 otherwise.

2. **Inter-Event Intervals**: time interval $W_k$ between two consecutive events $k-1$ and $k$
$$W_k = T_k - T_{k-1} \geq 0 \tag{3.17}$$

Therefore, the spiking activity of a neuron can be efficiently modeled using a point-process framework evaluating the Inter-Event Intervals. In particular, the ISI statistical distribution can be modeled as renewal process following well know probability models.

#### Spike train descriptors

A subset of statistical endpoints has been used to characterize the dynamic of the ISIs. [26] [22] The ISI of a certain neuron is defined as the set of $\mathbf{W} = [W_1, W_2, ...W_k, ...W_N]$ with N number of spikes identified for a recording signal of $T$ seconds. In particular, the following parameters have been employed:

- **Median**: $median = \frac{W_{N/2}+ISI_{N/2+1}}{2}$

- **Standard deviation**: $\sigma = \sqrt{E[(W-\mu)^2]}$
- **Coefficient of variation**: $CV = \frac{\sigma}{\mu}$
- **Number of spikes**: $N$
- **Skewness**: $\mu_3 = E[(\frac{W-\mu}{\sigma})^3]$
- **Kurtosis**: $\mu_4 = E[(\frac{W-\mu}{\sigma})^4]$
- **Frequency rate**: N/T [Hz]

### 3.4.2 Probality Density Function - PDF

Furthermore, the probability of an event to occur can be derived from the ISI statistical distribution based on the spike history. In literature, standard families of statistical distributions have been proven to suitably achieve a good fitting with ISI data.[5][27]

In this work, the distributions Inverse Gaussian, Gamma, Exponential, and Log Normal will be tested.

#### Inverse Gaussian

The electrical behaviour of a neuron cell can be mathematically described by the integrate-and-fire model, as in function 3.18

$$V(t) = V_0 + \beta t + \int_0^t dW(u)du \tag{3.18}$$

where V represents the potential across the cell membrane, $\beta$ is the drift parameter, and W(u) is a Wiener process (Brownian motion). Starting from the definition of this model, the equation of an Inverse Gaussian distribution can be analytically obtained.

The neural activity may indeed be modeled as an Inverse Gaussian distribution, which is considered a good estimation of the spiking events in time for a neural cell. [28][5] Therefore, it is possible to define the Probability

Density Function (PDF) of having a spike for a certain neuron identifying two parameters, as in equation

$$f(x|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda}{2x}\left(\frac{x-\mu}{\mu}\right)^2\right\} \tag{3.19}$$

with $\mu$ representing the mean of the distribution, while $\lambda = \frac{\mu^3}{\sigma^2}$ is the shape parameter, with $\sigma^2$ variance of the distribution.

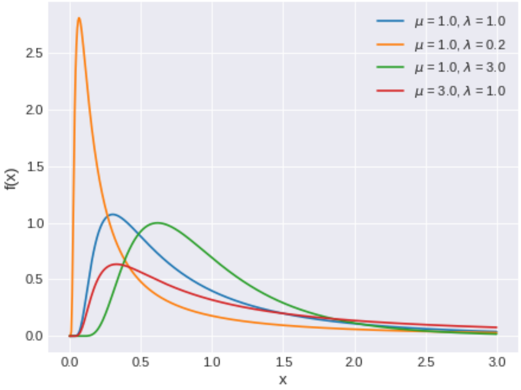

**Figure 3.4:** Inverse Gaussian distribution.

### Gamma distribution

The Gamma function is an extension of the factorial function and it is defined as

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx \tag{3.20}$$

The Gamma distribution is characterized by two parameters: a shape parameter $\alpha$ and a rate parameter $\beta$, both strictly positive, such that

$$\alpha = \frac{\mu^2}{\sigma^2}, \quad \beta = \frac{\mu}{\sigma^2} \tag{3.21}$$

with $\mu$ mean and $\sigma^2$ variance of the distribution. Stated that, it is possible to define the PDF of the Gamma distribution as

$$f(x|\alpha, \beta) = \frac{\beta^{\alpha} x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{3.22}$$

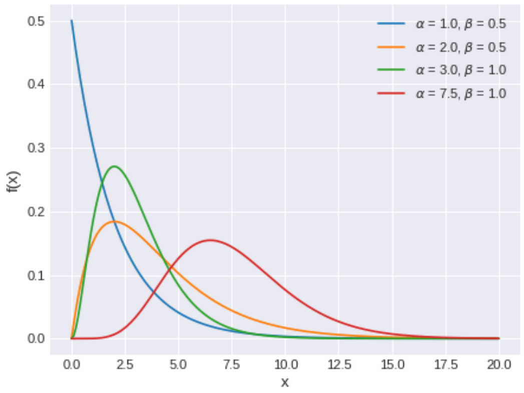

**Figure 3.5:** Gamma distribution.

### Exponential distribution

The Exponential distribution describes the time between events in a Poisson process. This is a particular case of the gamma distribution and it has the property to be memory-less. The parameter $\lambda$ that defines this distribution is called rate and the PDF is given by the function

$$f(x) = \lambda e^{-\lambda x} \tag{3.23}$$

### Log Normal distribution

The Log Normal distribution is a continuous probability distribution of a variable whose logarithm is normally distributes. It is defined with two parameters: $\mu$ referred to as location parameter, $\lambda$ as scale parameter such

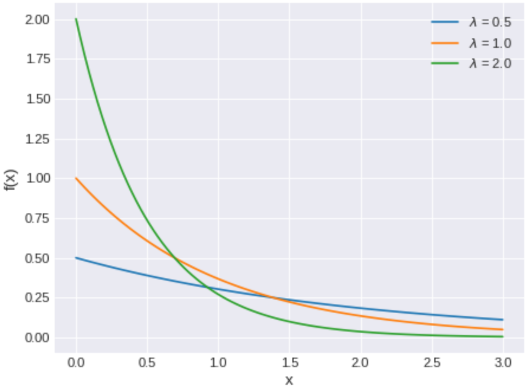

**Figure 3.6:** Exponential distribution.

that the mean is equal to $\{e^{\mu+1/(2\tau)}\}$ and the variance to $\{(e^{1/\tau}-1)e^{2\mu+1/\tau}\}$. Hence, the PDF is defined as

$$p(x|\mu,\tau) = \frac{1}{x}\sqrt{\frac{\tau}{2\pi}}\exp\left(-\frac{\tau}{2}(\ln x-\mu)^2\right) \tag{3.24}$$

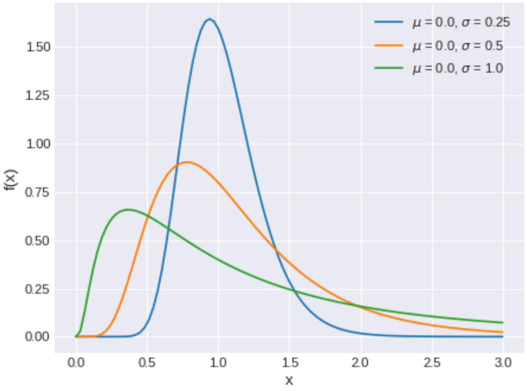

**Figure 3.7:** Log Normal distribution.

### 3.4.3 Point Process estimation

Given the vector of the parameters $\Theta$ defining the PDF and the ISI distribution $y$ of a neuron, it is possible to derive the value of the two parameters using the Bayes' theorem such that

$$p(\Theta|y) = \frac{p(y|\Theta)p(\Theta)}{p(y)} \tag{3.25}$$

Thus, the two parameters can be estimated by defining their prior distributions, and computing the ISI for each of th che identified neurons.

To this end, a Likelihood function is used to measure the goodness of fit of a certain model evaluated with the data. Maximizing such function the best combination ofthe set pf parameters $\Theta$ is obtained. Hence, the Maximum Likelihood Estimation (MLE) is employed and the results are tested by using the Kolmogorov-Smirnov test (KS-test). The KS-test is a non-parametric test that quantify the similarity of two continuous statistical distribution, the empirical one given by the data, and the theoretical one, defined by the model described by the identified parameters. In particular, the null hypothesis of this test is that the samples present the same distribution. To test this hypothesis, the Conditional Density Function (CDF) (equation 3.25) of the model $F_x(t)$ is compared to the one given by the empirical distribution.

$$F_x(t) = \int_0^t p(x|\Theta)d\Theta \tag{3.26}$$

#### Prior distribution

- Inverse Gaussian
  - $\mu_{IG} \sim \mathcal{U}(0.01, 0.1)$
  - $\lambda_{IG} \sim \mathcal{U}(0.01, 0.05)$
- Gamma
  - $\mu_\Gamma \sim \mathcal{U}(0.01, 0.1)$
  - $\sigma_\Gamma \sim \mathcal{U}(0.001, 0.5)$

- Exponential
  - $\lambda_{exp} \sim \mathcal{U}(1, 100)$
- Log Normal
  - $\mu_{LN} \sim \mathcal{U}(0.01, 0.1)$
  - $\tau_{LN} \sim \mathcal{U}(0.001, 1)$

### Posterior distribution

Given the prior distribution $p(\Theta)$, the posterior distribution could be estimated applying the Bayes' theorem. However, the marginal likelihood $p(y)$ needs to be assess in order to reach this goal. The Marginal Likelihood function defines the probability of the observed data $y$, in this case referring to the ISI of an identified neuron, as

$$p(y) = \int_{\Theta} p(y|\Theta)p(\Theta)d\Theta \tag{3.27}$$

However, this integral needs an high computational cost to be solved since it is evaluated in an high dimensional space of variables. Thus, it can be estimated rather than computed deterministically using an algorithm for sampling from a statistical distribution. In this work, the Hamiltonian Monte Carlo (HMC) Sampling algorithm will be used. [29]

Markov Chain Montecarlo methods are a group of statistical models aiming to obtain a sample of a desired distribution from the state of a Markov chain. This is a stochastic model of the state space that defines the state of a process after a number $n$ of steps given a transition probability matrix. If a sufficient history of the chain is known, using Monte Carlo estimation the equation 3.22 can be approximated as

$$\hat{p}(y) = \frac{1}{n}\sum_{i=1}^{n} p(y|\Theta^i) \tag{3.28}$$

with $\Theta^i$ equal to a set of parameter drawn by the a-priori distribution $p(\Theta)$. However, this approach is affected by the curse of dimensionality that yields to highly auto correlated steps and higher time of convergence when considering a large space of variables. By employing the HMC Sampling algorithm this issue is avoided because this method reduces the correlation between successive samples by moving to distant states that allows to maintain an high probability of acceptance.

## 3.5 Classification

Thus, the parameters defined in the previous paragraphs have been employed to build classification models able to characterize whether a neuron carries the LRRK2 mutation. In particular, first a binary classification during of the baseline dataset was performed, finally a 4-class classification has been run on the data collected after 24 hours.

### 3.5.1 Binary classification

The binary classification aims to discriminate between control neurons and LRRK2-mutated cells at the baseline phase. To this end, multivariate data analysis will be performed using both the spike train descriptors introduced in paragraph 3.5.1 and the results obtained applying the Point Process framework. First, the features have been normalized computing the z-score such as

$$z_i = \frac{x_i - \mu}{\sigma} \tag{3.29}$$

where $x$ is the *i-th* feature transformed into $z$ having mean equal to 0 ($\mu_z = 0$) and unitary standard deviation ($\sigma_z = 1$). Therefore, the dataset is split in a training set and a test set with a proportion of 75/25 stratified by outcome. The test set is not used in the training phase of the model in order to compute the accurateness of it preventing over-fitting problems. The following Machine Learning (ML) algorithms have been used to assess which one is more efficient in terms of results.

### Decision Tree Classifier

Decision Tree (DT) Classifier is a non-parametric model employed in supervised learning classification tasks. It aims to obtain simple classification rules to classify the inputs by maximizing the amount of information gain. Given a training subset **x** and a target vector **y**, a DT classifier having **n** nodes and **l** leaves is defined through a top-down greedy procedure. The expected information $I_{exp}$ needed to classify the two classes is defined at the node $n$ as

$$I_{exp}(n) = -\sum_{i=1}^{m} p_i log_2(p_i) \tag{3.30}$$

where $m$ corresponds to the number of possible classes and $p_i$ is the probability of an observation to belong to class $i$. In order to define the splitting rule at each node, the amount of Information explained by using each of the features is evaluated as

$$I(n|a) = -\sum_{i=1}^{m} p_i(a) log_2(p_i(a)) \tag{3.31}$$

and the best feature $a$ is found such that the Information Gain $I_G$ is maximized, such that

$$\max_a I_G = I_{exp} - I(n|a) \tag{3.32}$$

The algorithm is iterated at each node until the stopping rule is reached (minimum number of samples per leaf).

### Random Forest

Random Forest (RF) is a . RF algorithm builds a number of decision trees on bootstrapped training samples by selecting randomly a subset of predictors as split candidates from the full set of predictors. This means that a number of $n$ DTs are evaluated by using a bootstrap approach. This solution efficiently prevent over-fitting and high variance that DTs can experience.

### Support Vector Machines

Support Vector Machines (SVM) algorithm is a parametric supervised learning classifier aiming to estimate a set of hyper-planes that discriminate the samples in the features space. The equation of a generic linear hyper-plane is given by

$$\mathbf{x}_i^T \mathbf{w} + b = 0 \tag{3.33}$$

where $\mathbf{x_i}$ is the vector of the i-samples and $\mathbf{w}$ the coefficient vector for each feature. SVM aims to estimate the coefficient vector by maximizing the margin $2/||w||$ between the samples of the classes in the feature space such that

$$\begin{aligned} &\min_{w,b} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{m} \zeta_i \\ &s.t.\;\; y_i(\mathbf{x}_i^T \mathbf{w} + b) \geq 1 - \zeta_i \;\text{ with }\; \zeta_i \geq 0 \end{aligned} \tag{3.34}$$

where $\zeta$ and $C$ are regularization parameters used to address the case of non-fully separable classes. The SVM algorithm assumes that the data are linearly separable, but it may not exist an linear hyper-plane that achieve this goal. In this case, a Kernel function $K(x, x')$ can be applied to improve the classification results a linear boundary may be not sufficient. Thus, to reach a proper fitting the following types of kernel are used:

- Linear: $K(x, x') = \langle x, x' \rangle$
- Polynomial (of order $d$): $K(x, x') = (\gamma \langle x, x' \rangle)^d$
- Radial Based Function: $K(x, x') = \exp(-\gamma ||x - x'||^2)$
- Sigmoid: $K(x, x') = \tanh(\gamma \langle x, x' \rangle)$

with $\gamma$ regularization parameter that control the influence of a single training example.

### Logistic Regression

Logistic Regression (LR) is a linear classification model employed in supervised learning tasks. The algorithm create a model of the probability to have a certain outcome through a logistic function $\sigma(x)$ such that

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.35}$$

The penalty regularization functions employed are Lasso $L_1$, Ridge $L_2$, and Elastic-Net regularization, defined as follow:

- Lasso regularization ($L_1$)

$$\min_{w,b} ||\mathbf{w}|| + C \sum_{i=1}^{n} \log(\exp(-y_i(\mathbf{x_i}^T \mathbf{w} + b)) + 1) \tag{3.36}$$

- Ridge regularization ($L_2$)

$$\min_{w,b} \frac{1}{2}\mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{n} \log(\exp(-y_i(\mathbf{x_i}^T \mathbf{w} + b)) + 1) \tag{3.37}$$

- Elastic-Net regularization, which includes both $L_1$ and $L_2$ weighted by the parameter $\rho$

$$\min_{w,b} \frac{1-\rho}{2}\mathbf{w}^T \mathbf{w} + \rho||\mathbf{w}|| + C \sum_{i=1}^{n} \log(\exp(-y_i(\mathbf{x_i}^T \mathbf{w} + b)) + 1) \tag{3.38}$$

### 3.5.2 4-classes classification

A second multivariate analysis has been performed with the goal to classify four classes of neurons in the *After Stimulation* phase. In particular, the purpose is to verify if the statistical parameters computed are well suited to discriminate control and LRRK2-mutated neurons after 24 hours from the KA stimulation and how their behaviour changes compered to not stimulated networks. Thus, the four classes corresponds to:

- Control neurons stimulated with KA

- Control neurons not stimulated with KA
- LRKK2-mutated neurons stimulated with KA
- LRKK2-mutated neurons not stimulated with KA

First, a one-way ANOVA test [30] is performed to analyze whether the features are statistically significant different among the groups. To this end, three assumptions about the data need to be verified in order to perform the test:

- Independence of the samples.
- Population normally distributed.
- Population standard deviation is equal among the groups.

If these assumptioned are not satisfied, the Kruskal-Wallis H-test is employed. Then, the previous mentioned ML models have been employed. SVM is usually not well suited for multi-class classification, thus a strategy One versus Rest has been used to overcome this deficiency.

### 3.5.3 Hyper-parameters estimation

In order to achieve the best results in Machine Learning tasks, the regularization parameters of the model need to be set such that it can describe the behaviour of the training data reducing the risk of overfitting. The set of these parameters are referred to as hyper-parameters, which can be estimated through either of one of the following approaches:

- **Greedy approach**: every combination of parameters are considered in the parameter's space. The best combination is selected according to a specific metric by choice.
- **Bayesian approach**: the global maximum in the hyper-parameter space is searched starting from the hyper-parameters distributions. This is achieved iterating the maximum research creating "surrogate" models until the number of maximum iterations is reached.

However, a greedy-approach yields to a high computational complexity when the number of hyper-parameters is very high. In fact, given $k$ hyper-parameters and $n$ possible values that $k$ may assume, the total combinations to be assessed becomes equal to $\frac{(n+k-1)!}{k!(n-1)!}$. Therefore, the optimization algorithm Tree-structure Parzen Estimator (TPE) has been employed for the optimal hyper-parameters estimation. This method allows a low-complexity identification minimizing a loss function $f$. Given a hyper-parameter space $X$, the goal of this procedure is to obtain the optimal set of hyper-parameter $x_{opt}$ that minimizes the value of the loss function.

$$x_{opt} = \arg\min_{x \in X} f(x) \tag{3.39}$$

A Bayesian reasoning is used to construct the surrogate model and the best hyperparameters are selected iteratively using Expected Improvement, which is defined as the function:

$$\begin{aligned} EI_f(x) &= \int_{\infty}^{\infty} u(x) p_M(y|x) dy \\ u(x) &= \max(0, f' - f(x)) \end{aligned} \tag{3.40}$$

where $u(x)$ represents the utility function given by the maximum value between the current minimum obtained by $f(x)$ and the minimal value of the previous iterations $f'$. $EI_f(x)$ is evaluated on the whole hyper-parameters space for the model $M$ considering a threshold $y$ on the loss function. Therefore, this procedure yields to an estimation of the probability $p(x|y)$ by creating a generative process having a prior distribution of the hyper-parameters replaced with non-parametric densities such that

$$\begin{aligned} p(x|y) &= l(x) \quad if\ y < y^* \\ p(x|y) &= g(x) \quad if\ y \geq y^* \end{aligned} \tag{3.41}$$

with $l(x)$ density given by the observations with a loss function below the threshold $y^*$, $g(x)$ obtained from the rest of samples. Hence, at each iteration of the algorithm $l(x)$ will be a better estimation of the global minimum of the loss function. Furthermore, to prevent the overfitting issues, a 4-folds cross-validation is employed for the estimation of the loss function, such that

$$f(x) = 1 - \frac{1}{4}\sum_{k=1}^{4} a_k \tag{3.42}$$

where $a_k$ is the accuracy measure obtained for the *k-th* fold.

### 3.5.4 Classification assessment

Finally, the performances of classification are evaluated. To this end, a series of metrics are computed. Defining $N$ the number of samples, TP the True Positive samples, TN the True Negative, FP the False Positive, and FN teh False Negative it is possible to defined the following metrics:

- Accuracy $= \frac{TP+TN}{N}$
- Precision $= \frac{TP}{TP+FP}$
- Recall $= \frac{TP}{TP+FN}$

Moreover, the Receiver Operating Characteristics (ROC) curve can be build to assess the efficiency of the model. Hence, an additional metric is the area under the ROC curve, referred to as AUC ROC. The results will displayed in a confusion matrix. [31]

## 3.6 Computational tools

The code implemented to achieve the goals of this work is fully available in the following GitHub page. The code is implemented in a Python environment and it was run on Google Colaboratory Pro, given the high computational cost of most of the main functions.

#### Signal processing and spike sorting

For this section, the class `MEApy.py` was implemented. It comprises the methods needed to achieve the signal processing, for which was used the

SciPy library. A modified version of the Pywt library, which can be found in `https://github.com/SalvoCas/pywt` was employed for the spike detection through the CWT. Finally, the Sklearn library was employed to achieve the PCA and the clustering step.

### Point Process

The Python library PyMC3[32] has been employed for the estimation of the parameters of the Point Process framework. This lybrary is a Probabilistic Programming framework that allows to use predefined tools for the optimization of the fitting of complex models. The methods implemented are included in the class `PointProcess.py`.

### Classification

Finally, for the classification task the Machine Learning models described in this section were implementes by using the Python libraries Sklearn and Theano.

# Chapter 4

# Results

## 4.1 Spike sorting

### 4.1.1 Filtering

First, the raw signals have been filtered using the band-pass filter described in paragraph 3.3.1. Figure 4.1 shows the results of this step: the low frequency components, which are clearly visible in the raw data, are erased along with the background noise given by high frequency. Furthermore, the frequency component of raw signal has been studied to assess the presence of artifacts. Inspecting the reference channel by computing the PSD (figure 4.2), it is possible to notice two peaks representing noise artifacts at 439.45 Hz and its harmonic 878.90 Hz. This has been found on the whole dataset and may be due to instrumental distortion. Hence, a notch filter has been implemented to denoise the signal at those specific frequencies. The obtained time series have been used as input of the spike sorting pipeline.

### 4.1.2 Correlated channels

Hence, the pairwise correlation between the channels is assessed for each recording. An example is shown in figure 4.3. It is visible how specific channels carry the same information (dark blue squares). thus having a redundancy that needs to be addressed.

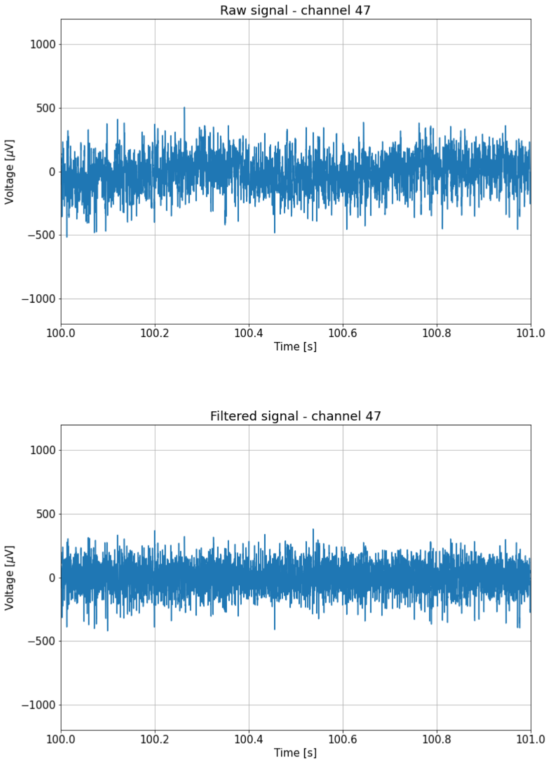

**Figure 4.1:** Signal of channel 47 depicted between 100 and 101 seconds before and after the filtering step. [Top] Raw signal. [Bottom] Filtered signal.

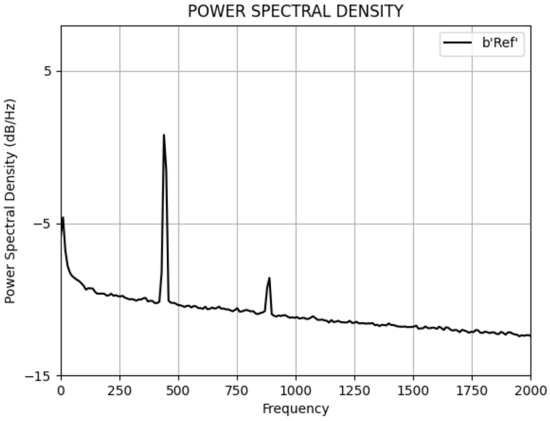

**Figure 4.2:** Power Spectral Density of reference channel

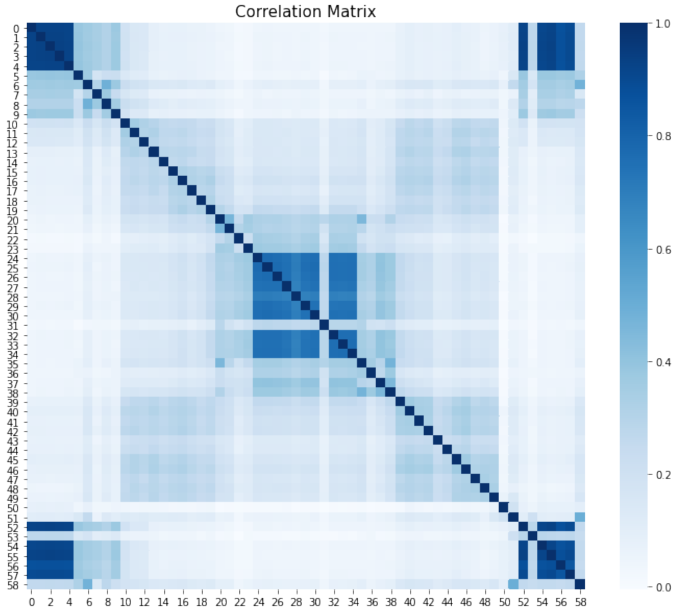

**Figure 4.3:** Pairwise correlation between the channels.

### 4.1.3 Spike detection

The CWT-based spike detection algorithm has been applied to detect the spike events for each signal. As introduced in the Material and Methods chapter, the algorithm defined by Nenadic[10] was implemented on a Python environment. This approach defined a new code pipeline for the characterization of the extracellular neural activity acquired by using Micro Electrode Arrays.

At this regard, the obtained results are compared to the Hard Threshold (HT) method used in the previous work, in order to outline the advantages of this new approach.

#### Estimation of parameters for CWT

First, it has been explored the optimal value of the parameter to be chosen for running the CWT-based spike detection algorithm. The results shown in this section are obtained using a omission cost parameter $L = -0.025$, the wavelet $($bior1.5), and a number of scales equal to 15 $(J = 15)$. These parameters have been chosen after different empirical trials, in order to optimize the balance between omissions and false detection.

In particular, in order to decide the value of the parameter $L$ two main factors has been taken into consideration. First, the signals having a dual dynamic, i.e. both negative and positive peaks, have been assessed with the Hard Threshold method. Thus, the spike detection algorithm has been run with different values of L. The results are depicted in figure 4.4. It is possible to notice that for negative values of L the two neurons are identifies as one due to the high number of false detection, whereas for positive values of L one of the two neuron is not detected due to the high number of omissions.

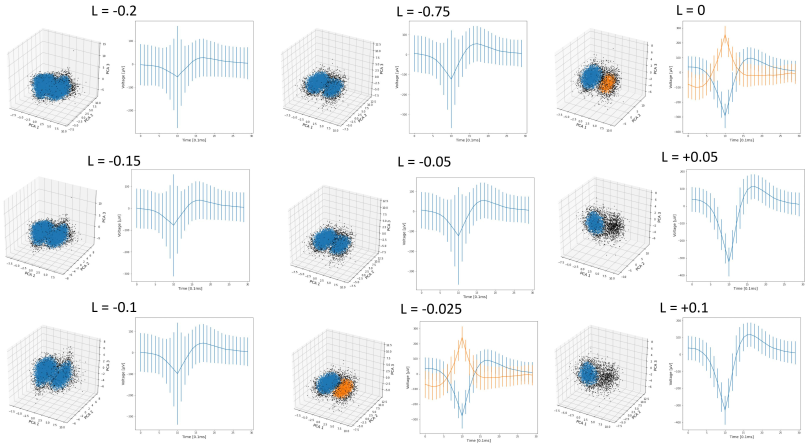

**Figure 4.4:** Clustering of spikes detected for different values of L.

Furthermore, a range of L around 0, approximately [-0.04; 0.02], allowed generally to obtain the best results in terms of statistical neural modelling.

### Comparison between CWT and HT approaches

Figure 4.5 depicts the filtered signal with the detected spikes (red dots). The procedure takes into account a refractory period of the neuron equal to 3 ms. It is possible to compare the results obtained with the ones of the previous work,[13] in which a hard threshold-based (HT) approach was used. Considering all the channels of the recording `2018-11-27T10-40-53POP3BL.h5`. The following images will take into account channel b'37', and the spikes detected using the CWT-based, and the Hard Threshold (HT) detection

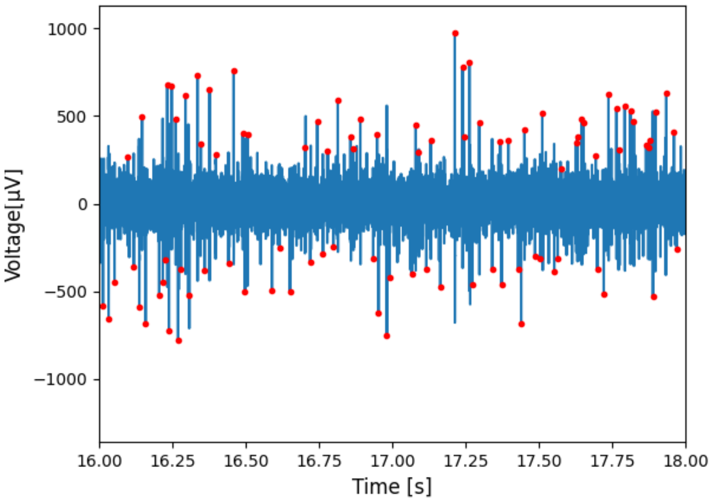

**Figure 4.5:** Portion of signal (blue) with detected spike instants (red dots).

algorithm with a threshold equal to $\pm 4 * MAD$.

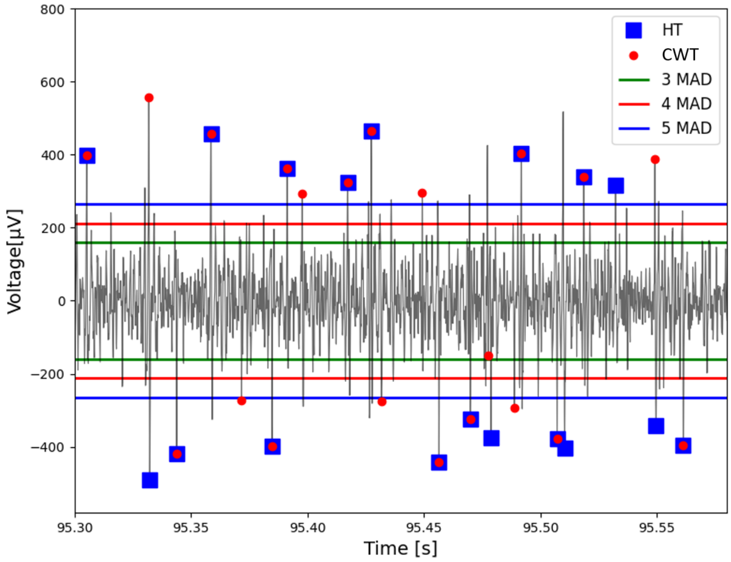

**Figure 4.6:** Comparison between spikes detected using an hard threshold approach (HT, blue squares), and spikes detected by the CWT-based algorithm (red dots). The thresholds corresponding to different levels of MAD are shown.

In figure 4.6 a portion of the signal is shown, with the spikes detected by using both methods outlined (red dots: CWT, blue squares: HT), along with the MAD thresholds at different levels (3, 4, 5 times). It is possible to notice how some of the detected spikes overlap, whereas the CWT-based approach is able to detect events having a peak near or below the MAD thresholds. Thus, the waveform of each spike has been extracted and PCA has been run.

### 4.1.4 Clustering

The DBSCAN clustering algorithm has been applied to identify the neurons, detecting clusters of spikes in the PCA space. The parameters chosen are $\varepsilon = 1.1$ and $MinPts = 80$. In figure 4.7 a three-dimensional plot of detected spikes for a channel is depicted in the space defined by the first three principal components. The spikes are assigned to either the orange or the blue neuron, or identified as noise (black dots). The right portion of the image shows the mean waveform of the two neurons, that have a different dynamic in terms of peak (positive-negative) and shape. The DBSCAN clustering algorithm is effectively able to detect clusters within the lower-dimensional dataset obtained after the PCA, classifying as outliers those spikes with no similar characteristics with others.

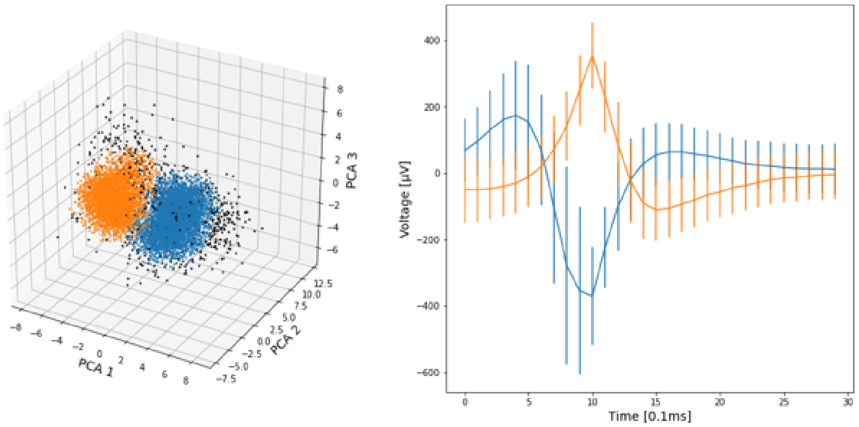

**Figure 4.7:** [Left] Detected spikes depicted according to their first three principal components. The neurons are clustered (blue and orange cluster), and outliers (black points) identified as noise. [Right] Mean waveform of each detected neuron.

## 4.2 Modelling of neurons

After the spike sorting procedure, the Inter Spike Interval (ISI) of each neuron has been computed. In figure 4.8 the Inter Spike Interval (ISI) of the neuron across time is shown, along with the 200 samples moving average (red line). Thus, by inspecting the trend of the signal, it was possible to notice that the ISI is stationary after 100 seconds. Moreover, the signals have different recording time, but always greater than 300 seconds. Hence, the interval between 100 s and 300 s has been chosen to inspect the spike trains in the further analysis for stationary purpose and to maintain the analysis consistent across the recordings.

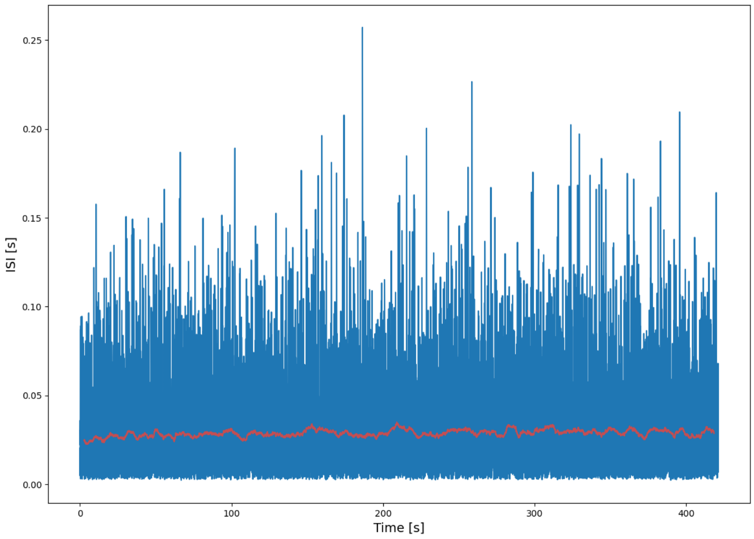

**Figure 4.8:** Inter Spike Interval across time. In red the moving average for 200 samples.

The Point Process framework aims to model the probability distribution to have a consecutive event (the spike) in an interval of time $\Delta$. The statistical distribution introduced in section 3.4.2 have been tested and the results compared in order to evaluate the goodness of fit. In figure 4.9 the ISI histogram is depicted with the fitted Inverse Gaussian PDF superimposed. By performing the Kolgomorov-Smirnov test the model is able to fit the empirical data.

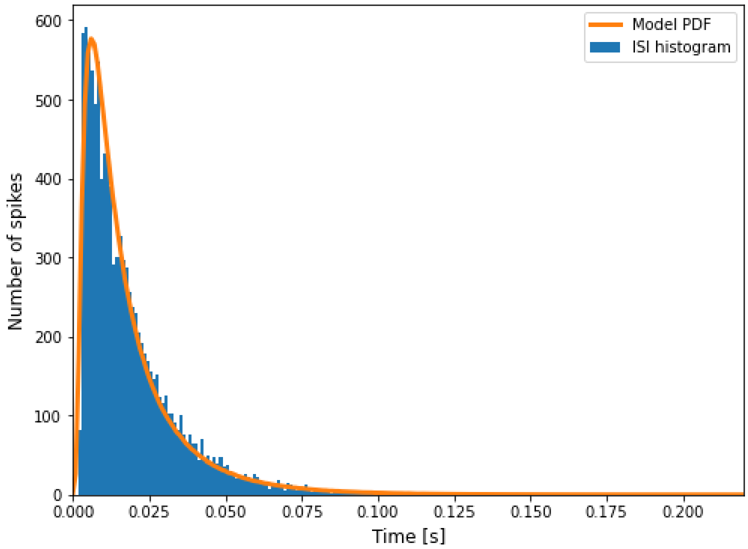

**Figure 4.9:** Inverse Gaussian fitted PDF superimposed to the ISI histogram of the neural data

By inspecting the KS plots it possible to notice that the Inverse Gaussian model remains inside the significant boundaries defined by the Kolgomorov-Smirnov test, i.e. the model is statistically significant with a p-value $<$ 0.05. However, this does not happen if the other models are employed. In fact, none of the other model is able to fit the data. This is visible by inspecting the KS plot obtained fitting the empirical data using the mentioned distribution (figure 4.11); in fact, the models do not remain inside the significance boundaries.

Thus, the Inverse Gaussian model was used for this task and the parameter estimated as shown in section 3.4.3. In particular, the HMC algorithm has been used and 12 Markov Chains have been deployed. In order to achieve a convergence of the algorithm, the neurons having less than 1000 spikes in the interval between 100 s and 300 s have been excluded and they are not considered in the following analysis. The convergence of the algorithm for the two parameters of the Inverse Gaussian is shown in figure 4.12.

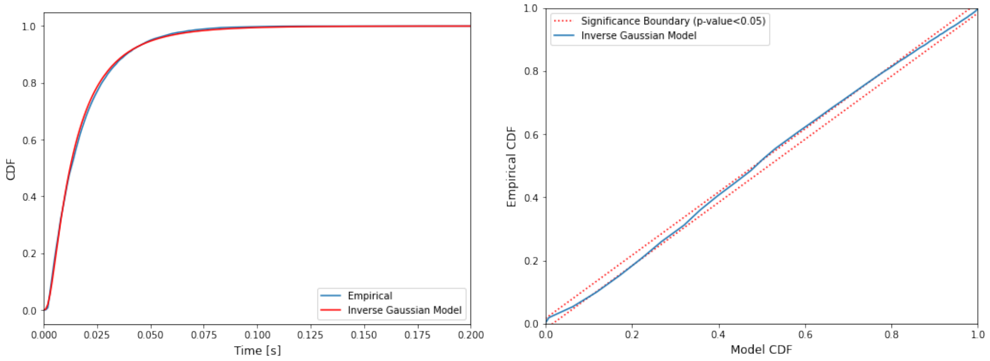

**Figure 4.10:** Inverse Gaussian model - KS plots. [Left] Comparison between empiral CDF (blue) and model CDF (orange). [Right] KS test on the model, in orange the significance boundaries (p-value<0.05).

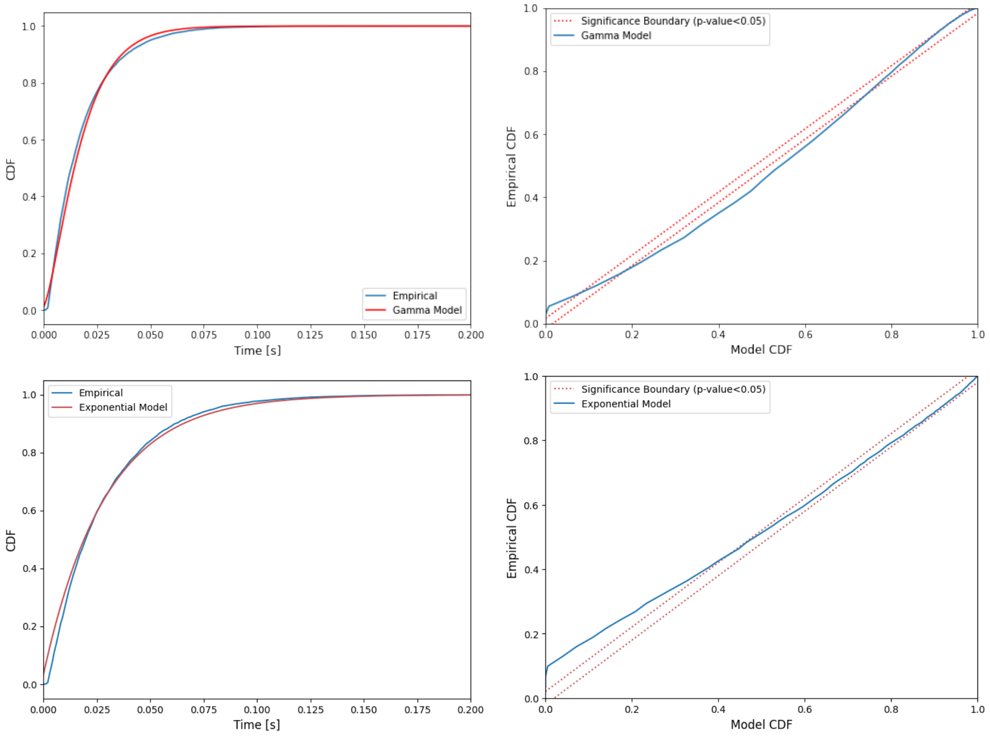

**Figure 4.11:** [Top] KS plots for Gamma Distribution model. [Bottom] KS plots for Exponential Distribution model.

### 4.2.1 Data preparation

After the spike sorting phase, the identified neurons that satisfy the conditions imposed to apply the point process are divided as in table 4.1. A

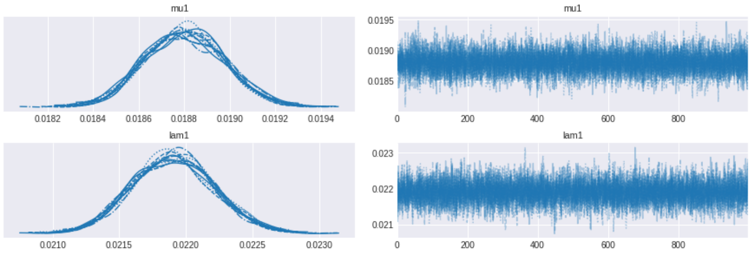

**Figure 4.12:** Trace plot of the Point Process features. Each line represents a Markov Chain.

total number of 531 neurons are identified in the *Baseline* phase, whereas 458 in the *After 24 hours* phase, subdivided in control neurons (201) and neurons stimulated through Kainic acid (257). It is possible to notice that the number of LKKR2-mutated neurons identified after the KA stimulation is much smaller than the healthy ones. The same trend is visible for the baseline recording.

| | *Baseline* | *After 24 hours* | | Total |
|---|---|---|---|---|
| | | Control | KA stimulated | |
| Healthy | 334 | 101 | 177 | 612 |
| LRRK2 | 197 | 91 | 80 | 368 |
| Total | 531 | 201 | 257 | 980 |

**Table 4.1:** Number of detected neurons.

The results obtained with the Point Process framework are shown in figure 4.13. The statistical parameters described in section 3.4.1 are computed and, along with the Point Process features, and they are employed to perform the classification tasks (figure 4.14).

Moreover, in Table 4.2 and 4.3 the mean firing rate of the two population are depicted respectively for the *Baseline* and *After 24 hours* phase.

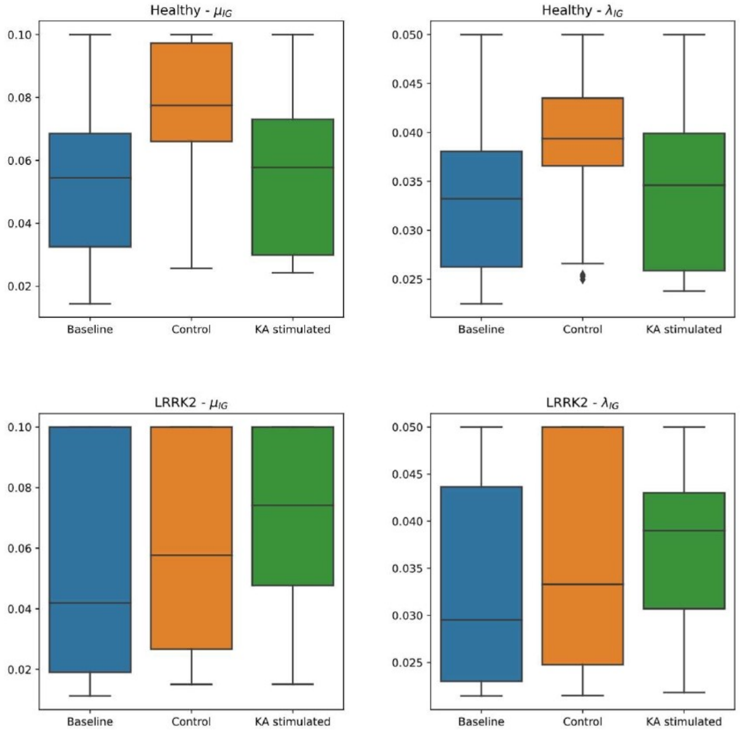

**Figure 4.13:** Boxplot of extracted features of the Inverse Gaussian model divided by group affiliation.

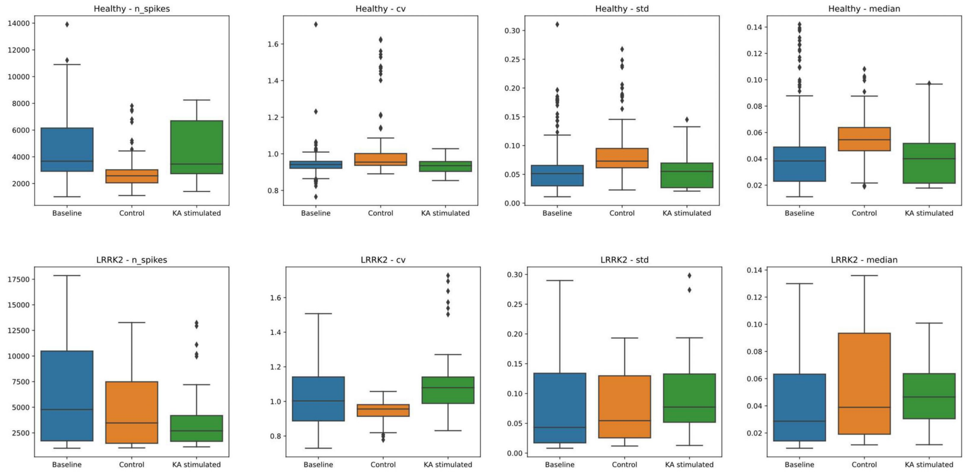

**Figure 4.14:** Spike train statistic descriptors.

| Population | Firing rate [Hz] |
|---|---|
| Healthy neurons | 22.97 ± 12.33 |
| LRRK2-mutated neurons | 31.90 ± 25.09 |

**Table 4.2:** Firing rate at *Baseline* phase (mean±std).

| Population | Control | KA stimulated |
|---|---|---|
| Healthy neurons | 14.07 [Hz] ± 6.55 | 22.55 [Hz] ± 10.86 |
| LRRK2-mutated neurons | 25.11 [Hz] ± 17.60 | 19.05 [Hz] ± 12.92 |

**Table 4.3:** Firing rate at *After 24 hours* phase (mean±std)

## 4.3 Classification

### 4.3.1 Binary classification

First, a binary classification aiming to distinguish healthy and mutated cells at the Baseline phase has been performed. In order to further evaluate the results, the models have been tested using both standardized features and the results of PCA.

In Table 4.4 the metrics obtained for each ML model are shown. Overall, the PCA does not improve the classification results since the metrics already reach high values. The best performance is obtained by applying the Random

| ML model | Accuracy | Recall | Precision | ROC AUC |
|---|---|---|---|---|
| DT | 95.38(93.85) | 93.33(86.67) | 93.33(95.12) | 94.90(92.16) |
| LR | 93.85(93.85) | 86.67(86.67) | 95.12(95.12) | 92.16(92.16) |
| SVM | 94.62(94.62) | 93.33(93.33) | 91.30(91.30) | 94.31(94.31) |
| RF | 96.15(93.08) | 95.56(91.11) | 93.48(89.13) | 96.01(92.61) |

**Table 4.4:** Results of *Baseline* classification. Between brackets results obtained with PCA.

Forest Classifier with an area under the ROC curve of the 96.01. The confusion matrix in figure 4.15 depicts the results obtained for the RF. Figure 4.16 shows the ROC curves of the models compared to a random classifier.

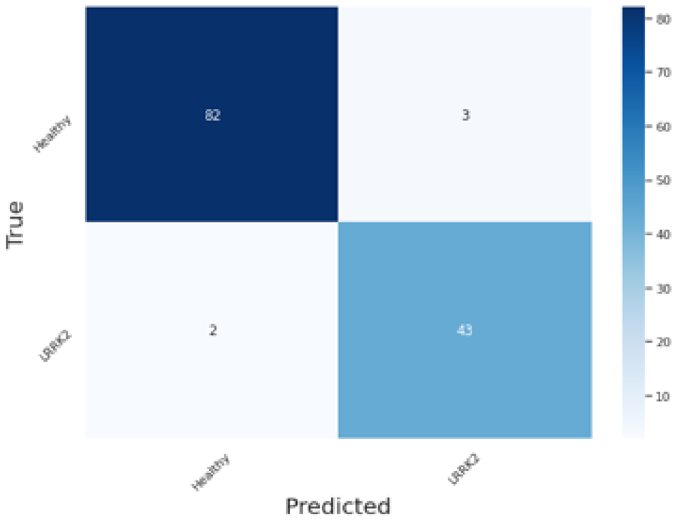

**Figure 4.15:** Confusion matrix for baseline classification applying the Random Forest model.

| ML model | Accuracy | Recall | Precision |
|---|---|---|---|
| DT | 53.98(54.87) | 53.98(54.87) | 60.92(53.23) |
| LR | 55.75(55.75) | 55.75(55.75) | 53.14(53.14) |
| SVM | 64.60(66.37) | 63.60(66.37) | 63.60(66.42) |
| RF | 74.80(74.80) | 74.80(74.80) | 77.82(77.82) |

**Table 4.5:** Results of *After 24 hours* classification. Between brackets results obtained with PCA.

### 4.3.2 After 24 hours classification

The features have been tested using a one way ANOVA to study the statistical significance of the difference among groups. The null hypothesis

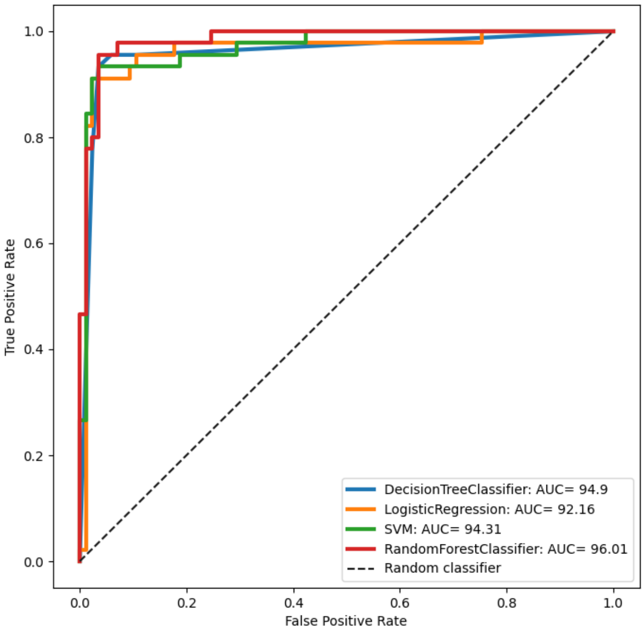

**Figure 4.16:** ROC curves for *Baseline* phase

was accepted for every considered features (p-value $< 0.001$), showing a statistical significant difference between the groups.

Then, a 4 class multivariate classification has been performed. The results are shown in table 4.5. Furthermore, the results obtained applying the Random Forest classifier are shown in figure 4.17. Also for this task RF model resulted having the best performances, reaching an accuracy of 74.8 and a precision of 77.82. The application of PCA did not improve the results of the models.

It is possible to notice a good classification for the stimulated populations, well defined by the model. However, a low degree of accuracy is observed for the classification of healthy control networks, with 17 misclassifications out of 25.

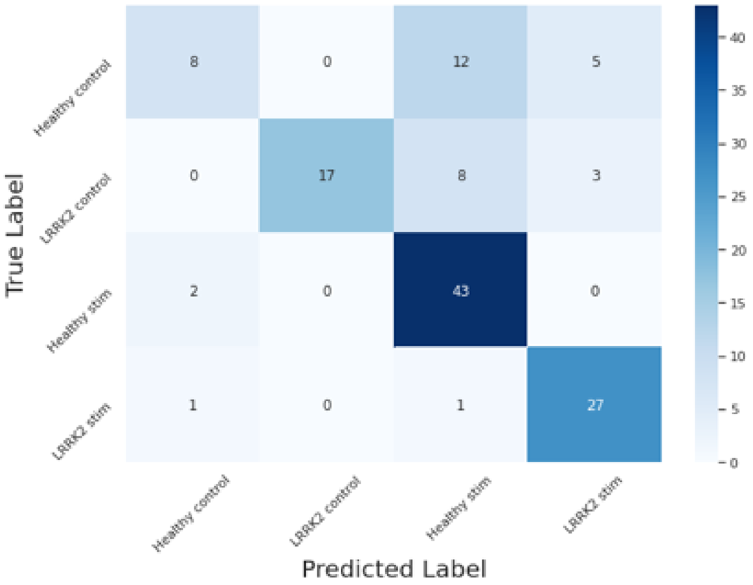

**Figure 4.17:** Confusion matrix for after 24 hours multi-class classification applying the Random Forest model.

# Chapter 5

# Conclusion

The presented work is aimed at improving the statistical characterization of the spiking activity recorded from *in-vitro* neural networks carrying the mutation G2019S on the gene LRRK2. This study stands as continuation of a previous work developed at Politecnico di Milano with the collaboration of the Department of Neuromedicine and Movement Science of the Norwegian University of Science and Technology (NTNU) of Throdheim (Norway) that first has provided an analysis of the bioelectric behaviour of the dataset here considered. [13]

The main innovative characteristic of the study presented can be summarized as:

- An improvement in the spike detection step of the spike sorting pipeline introducing an algorithm based on Continuous Wavelet Transform rather than hard thresholds.
- The stochastic Inverse Gaussian distribution has been proven to correctly model the spiking behaviour of the identified neurons.
- The extracted features along with a series of statistical metrics derived from the Inter-Spike Interval distributions have improved the results of the classification both in the *Baseline* and *After 24 hours* phase.
- The obtained results act as a consolidation of previous performed analyses in terms of statistical characterization of the differences between

the activity of LRRK2-mutated neural networks compared to a control healthy population.

In the following paragraph, this improvements are discussed and the results compared to the state of the art.

### Discussion of the results

MEA has proven to be an important tool to study the spiking activity of neural cells. However, the signals acquired by the electrode are inevitably corrupted by noise, generated by many sources, most notably by the micro-electrode itself or the activity of distant neurons.

Thus, a reliable spike sorting pipeline is necessary. This work has shown that a spike detection approach based on Continuous Wavelet Transform is able to identify spikes having limited intensity. This overcomes one of the main issues depicted by previous analysis,[13] where the threshold methodology based on MAD was not able to consider non-stationarities in the recording.

In figure 5.1, the identification of the time instant of the events are depicted, along with the MAD thresholds assessed in the previous work. It is possible to notice that there are some overlapping identifications, however the CWT-based spike detection algorithm implemented efficiently recognizes events occurring with a voltage intensity lower than the statistical thresholds. Therefore, this step becomes more dynamical and it does not rely on the behaviour of the entire signal, rather than on the local trend. Furthermore, the employment of a biorthogonal mother wavelet (*bio1.5*), which resembles the typical shape of an action potential, at time scales of a typical spike event (i.e. between 0.5 ms and 2 ms), allows an identification based on *a-priori* information well know in literature, lowering the possibility to have false detections that would consider noisy components as spikes.

The DBSCAN clustering method is able to sort group of similar spikes, thus identifying the neural cells based on their bioelectric behaviour. Furthermore, this approach does not consider noisy spikes which are identified as outliers and filtered out. This method allows to improve further the quality of the analysis.

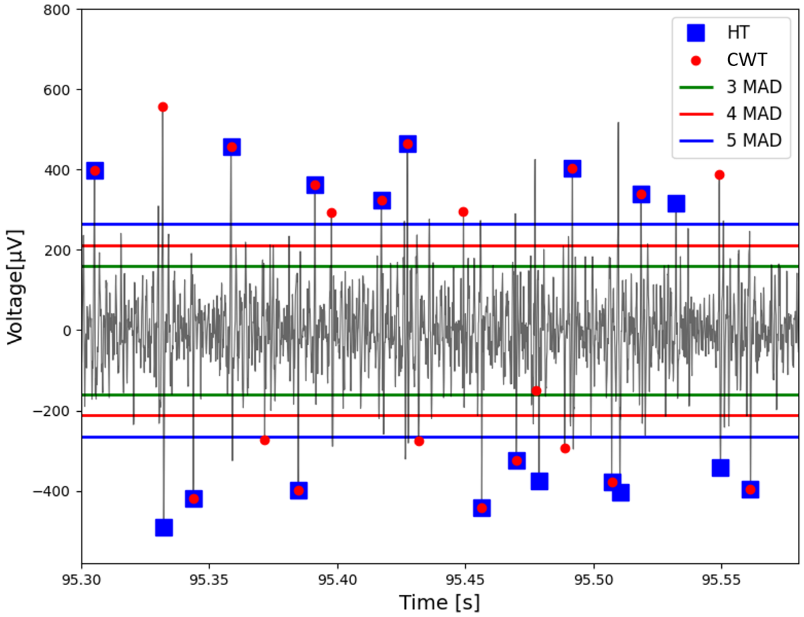

**Figure 5.1:** Comparison between spikes detected using an hard threshold approach(HT, blue squares), and spikes detected by the CWT-based algorithm(red dots). The thresholds corresponding to different levels of MAD are shown.

After performing the spike sorting technique, the Inter Spike Interval of the obtained neurons was assessed. The goodness-of-fit was tested for different state-of-art statistical distributions and it is found out that the detected neurons follow a typical Inverse Gaussian. In fact, the results of Kolgomorov-Smirnov test suggest the statistically significance of the model fitted on the empirical data. This result concurs with the literature on the spiking activity of *in-vitro* neural cells.[5]

The performances of the multi-variate analysis have shown considerable improvements if compared to the results previously obtained. The estimated features of the Point Process in addition to spike train descriptors (section 3.4.1) have allowed to increase the overall goodness of the classification for the *Baseline* phase for every Machine Learning model applied. (Table 2.1 and Table 4.2) Thus, it can be stated that the approach here presented better outlines differences between the two population during this phase. The same results are shown for the multiclass classification at he *After 24 hours* phase. In particular, the Random Forest classifier reaches an accuracy of 74.8 (previous 68.6).

During the *Baseline* phase, a greater activity has been observed for LRRK2-mutated cells compared to the healthy control networks. This trend has been not detected by the previous analysis [13], however it corroborates the results obtained by Valderahaug et al., 2015.[6] In fact, this study has confirmed an increase baseline network activity. This is correlated to a significant difference in mitochondrial content within the cells at this stage, observed by the previously mentioned research group. At this regard, several studies has linked the increment in spiking activity in mutated neural networks to a pathogenic change preceding dentritic alterations in cortical neurons.[33] In fact, the overexpression of the G2019S mutation of the gene LRRK2 increases basal synaptic efficiency through a postsynaptic mechanism, and disrupts long-term depression.[7][8] Thus, this results may identify a pathogenic function of this genetic mutation.

A different behaviour has been observed between the two populations after the neurotoxic insult. In fact, LRRK2-mutated cells presents a statistical significant decreasing in the spiking frequency for the stimulated networks. This trend can bee seen by evaluating the change in the parameters of the Point Process model. Previous works [6] have observed a drop in the spiking activity of LRRK2-mutated neural networks 24 hours after a overexcitation of the cells through Kainic acid. The results here presented (figure 5.2) corroborate this conclusions, since a considerable drop in the mean firing rate has been depicted comparing both baseline and control recordings to the ones acquired after the stimulation.

Kainic Acid is a strong agonist of the ionotropic transmembrane receptors for glutammate AMPA [24] which regulate fast synaptic transmission in the central nervous system. The results obtained may be explained with the hypothesis that LRRK2-mutated neurons are not able to overcome the insult, maintaining the damage rather than recovering the functioning mechanisms. On the contrary, the opposite is observed for healthy cells, meaning that an increment in the spiking activity may address the task to regenerate properly the correct functions of the neural cells.

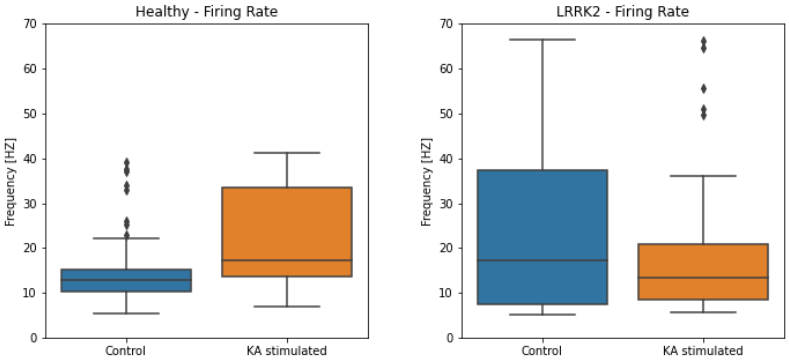

**Figure 5.2:** Firing rate [HZ] in the *After 24 hours* phase divided by groups

Furthermore, the results are consistent with the conclusions depicted by Valderhaug et al.[6], that have shown a decreasing mean fire rating for LRRK2 neural networks during KA stimulation compared to the control neural networks.

### Limitations

Several limitations to this work are present. The clustering method can be further improved by optimizing the input parameters of the DBSCAN algorithm to be specifically fitted for each one of the signal. This would allow a better identification of the neuron starting from the spike train. Moreover, different features extraction methods could be employed to improve the identification of the neurons. In fact, the structure of the data may be difficult to fully be delineated using only PCA. Studies have shown different approaches that can improve this aspect as mixed algorithms that consider a broad set of features.[34]

Furthermore, the current analysis could be expanded by exploring the burst activity within the network. In fact, bursting has been used as a key feature to characterize the electric behaviour of *in-vitro* neurons, playing an important role in the communication between the cells.[35]

Finally, the statistical representation of the neural behaviour obtained by using the Point Process framework provides static information about the characteristics of the cell activity. Therefore, a dynamical application of Point Process model can be used to understand how the system changes in time. To this end, a instant Point Process evaluation can be employed by estimating the time-varying parameters of the Inverse Gaussian.[36] This approach would enrich the quality of the information on the nature of the impact that the G2019S mutation has on the bioelectric activity of human neural cells by provide further insights into the evolution in time of the neural interconnections. This is particularly important to achieve a better characterization of the the response to a neurological insult, such as the Kainic acid stimulation, and the recovery from it.

# Bibliography

[1] R. van de Wijdeven, O. H. Ramstad, V. D. Valderhaug, P. Köllensperger, A. Sandvig, I. Sandvig, and Øyvind Halaas, “A novel lab-on-chip platform enabling axotomy and neuromodulation in a multi-nodal network,” *Biosensors and Bioelectronics*, vol. 140, p. 111329, 2019.

[2] K. Takahashi and S. Yamanaka, “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors,” *Cell*, vol. 126, no. 4, pp. 663–676, 2006.

[3] J.-Q. Li, L. Tan, and J.-T. Yu, “The role of the lrrk2 gene in parkinsonism,” *Molecular Neurodegeneration*, vol. 9, no. 47, 2014.

[4] S. Bardien, S. Lesage, A. Brice, and J. Carr, “Genetic characteristics of leucine-rich repeat kinase 2 (lrrk2) associated parkinson’s disease,” *Parkinsonism and related disorders*, vol. 17, no. 7, pp. 501–508, 2011.

[5] S. Iyengar and Q. Liao, “Modeling neural activity using the generalized inverse gaussian distribution,” *Biological Cybernetics*, vol. 77, pp. 289–295, 1997.

[6] V. D. Valderhaug, O. H. Ramstad, R. van de Wijdeven, K. Heiney, S. Nichele, A. Sandvig, and I. Sandvig, “Structural and functional alterations associated with the lrrk2 g2019s mutation revealed in structured human neural networks,” *BioRxiv*, 2020.

[7] E. S. Sweet, B. Saunier-Rebori, Z. Yue, and R. D. Blitzer, “The parkinson’s disease-associated mutation lrrk2-g2019s impairs synaptic plastic-

ity in mouse hippocampus," *Journal of Neuroscience*, vol. 35, no. 32, pp. 11190–11195, 2015.

[8] B. A. Matikainen-Ankney, N. Kezunovic, R. E. Mesias, Y. Tian, F. M. Williams, G. W. Huntley, and D. L. Benson, "Altered development of synapse structure and function in striatum caused by parkinson's disease-linked lrrk2–g2019s mutation," *Journal of Neuroscience*, vol. 36, no. 27, pp. 7128–7141, 2016.

[9] R. HG, P. C, and Q. Q. R., "Past, present and future of spike sorting techniques," *Brain Research Bulletin*, pp. 106–17, 2015.

[10] Z. Nenadic and W. B. Joel, "Spike detection using the continuous wavelet transform," *IEEE transactions on bio-medical engineering*, vol. 52, no. 1, pp. 74–87, 2005.

[11] A. Elbaz, L. Carcaillon, S. Kab, and F. Moisan, "Epidemiology of parkinson's disease," *Revue Neurologique*, vol. 172, no. 1, pp. 14–26, 2016. Neuroepidemiology.

[12] G. W. Huntley, P. Massobrio, J. Tessadori, M. Chiappalone, and M. Ghirardi, "In vitro studies of neuronal networks and synaptic plasticity in invertebrates and in mammals using multielectrode arrays," *Neural Plasticity*, 2015.

[13] R. Levi, "Characterization of spiking activity in structured in-vitro human neural networks through point process statistical modelling approaches," *Politecnico di Milano*, 2020.

[14] A. Salles, J. G. Bjaalie, K. Evers, M. Farisco, B. T. Fothergill, M. Guerrero, H. Maslen, J. Muller, T. Prescott, B. C. Stahl, H. Walter, K. Zilles, and K. Amunts, "The human brain project: Responsible brain research for the benefit of society," *Neuron*, vol. 101, no. 3, pp. 380–384, 2019.

[15] D. V. V. Essen and M. Glasser, "The human connectome project: Progress and prospects," *Cerebrum: the Dana Forum on Brain Science*, vol. 2016, 2016.

[16] C. Thomas, P. Springer, G. Loeb, Y. Berwald-Netter, and L. Okun, "A miniature microelectrode array to monitor the bioelectric activity of cultured cells," *Experimental Cell Research*, vol. 74, no. 1, pp. 61–66, 1972.

[17] S. G. Reich and J. M. Savitt, "Parkinson's disease," *Medical Clinics of North America*, vol. 103, no. 2, pp. 337–350, 2019. Neurology for the Non-Neurologist.

[18] H. Deng, P. Wang, and J. Jankovic, "The genetics of parkinson disease," *Ageing Research Reviews*, vol. 42, pp. 72–85, 2018.

[19] A. G. Henry, S. Aghamohammadzadeh, H. Samaroo, Y. Chen, K. Mou, E. Needle, and W. D. Hirst, "Pathogenic LRRK2 mutations, through increased kinase activity, produce enlarged lysosomes with reduced degradative capacity and increase ATP13A2 expression," *Human Molecular Genetics*, vol. 24, pp. 6013–6028, 08 2015.

[20] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of Physiology*, vol. 117, no. 4, pp. 500–544, 1952.

[21] R. Benitez and Z. Nenadic, "Robust unsupervised detection of action potentials with probabilistic models," *IEEE transactions on bio-medical engineering*, vol. 54, no. 4, pp. 1344–1354, 2008.

[22] A. M. Tukker, F. M. Wijnolts, A. de Groot, and R. H. Westerink, "Human ipsc-derived neuronal models for in vitro neurotoxicity assessment," *NeuroToxicology*, vol. 67, pp. 215–225, 2018.

[23] L. Chen, Y. Deng, W. Luo, Z. Wang, and S. Zeng, "Detection of bursts in neuronal spike trains by the mean inter-spike interval method," *Progress in Natural Science*, vol. 19, no. 2, pp. 229–235, 2009.

[24] J. A. Davies, "Kainic acid," in *xPharm: The Comprehensive Pharmacology Reference* (S. Enna and D. B. Bylund, eds.), pp. 1–3, New York: Elsevier, 2007.

[25] O. O. Anoh, R. A. Abd-Alhameed, S. M. Jones, J. M. Noras, Y. A. Dama, A. M. Altimimi, N. T. Ali, and M. S. Alkhambashi, "Comparison of orthogonal and biorthogonal wavelets for multicarrier systems," in *2013 8th IEEE Design and Test Symposium*, pp. 1–4, 2013.

[26] J. A. Bradley, H. H. Luithardt, M. R. Metea, and C. J. Strock, "In vitro screening for seizure liability using microelectrode array technology," *Toxicological Sciences*, vol. 163, no. 1, pp. 240–253, 2018.

[27] H. Konno and Y. Tamura, "Stochastic modeling for neural spiking events based on fractional superstatistical poisson process," *AIP Advances*, vol. 8, no. 1, p. 015118, 2018.

[28] D. Perkel, G. Gerstein, and M. GP, "Neuronal spike trains and stochastic point processes," *Biophysical Journal*, vol. 7.4, pp. 391–418, 1967.

[29] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid monte carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216–222, 1987.

[30] K. Moder, "Alternatives to f-test in one way anova in case of heterogeneity of variances (a simulation study)," *Psycological Test and Assesment Modeling*, vol. 52.4, pp. 343–353, 2010.

[31] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, vol. 17, no. 1, pp. 168–192, 2021.

[32] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in python using pymc3," *PeerJ Computer Science*, vol. 2, p. e55, 2016.

[33] E. Plowey, J. Johnson, E. Steer, W. Zhu, D. Eisenberg, N. Valentino, Y.-J. Liu, and C. Chu, "Mutant lrrk2 enhances glutamatergic synapse activity and evokes excitotoxic dendrite degeneration," *Biochimica et biophysica acta*, vol. 1842, 05 2014.

[34] R. Bestel, A. W. Daus, and C. Thielemann, "A novel automated spike sorting algorithm with adaptable feature extraction," *Journal of Neuroscience Methods*, vol. 211, no. 1, pp. 168–178, 2012.

[35] T. Ishii and T. Hosoya, "Interspike intervals within retinal spike bursts combinatorially encode multiple stimulus features," *PLOS Computational Biology*, vol. 16, pp. 1–30, 11 2020.

[36] R. Barbieri, E. C. Matten, A. A. Alabi, and E. N. Brown, "A point-process model of human heartbeat intervals: new definitions of heart rate and heart rate variability," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 288, no. 1, pp. H424–H435, 2005.