

Polytechnic University of Madrid

Engineering the future



Computer Systems Department

Higher Technical School of Computer Systems Engineering

EXPLORATION OF PARKINSON'S DISEASE RECOGNITION SPACE BY ARTIFICIAL NEURON NETWORKS.

Written by:

Mario Rojo Vicente

Directed by:

Francisco Díaz Pérez

Bachelor Thesis for the Software Engineering Degree

June 2022

Acknowledgment

Abstract

Summary

1	Introduction	4
1.1	Context	4
1.2	Objective	6
1.3	Document Structure	7
1.4	State Of Art	8
2	Data	9
2.1	Cleaning The Data Set	10
2.2	Normalizing The Data Set And Transforming Non Numerical Values . .	12
2.3	Splitting The Data Set Into Training, Test and Validation	15
2.4	Generating The Labels	16
2.5	Balancing The Data Set	17
3	Theoretical Bases	18
4	Conclusion	22
4.1	Social Implications	23
5	Possible Future Work Lines	24

1 Introduction

1.1 Context

The origin of Parkinson's disease is at best unclear. While some authors claim it to be a consequence of unknown toxins introduced in the XIX century during the Industrial Revolution, many historical evidences suggest that the presence of this ailment goes as far back as the 2000-1500 ac, when some of its symptoms were first referenced in the Vedas texts. Further references to what could possibly be Parkinson's disease were also found in Egyptian papyrus and ancient Chinese medical treaties, among other historical documents. Also, several historical personalities, such as Hipocrates, an ancient Greek's doctor, and the Italian genius Leonardo Da Vinci, analyzed and studied several symptoms related to this illness.

The first publication referencing what we currently know as Parkinson's disease was a paper by the name "An essay on the shaking palsy", published in 1817 by the British surgeon James Parkinson. None the less, the disease did not get its final name until the year 1880 when Jean-Marie Charcot, a well recognized french neurologist that was researching the stiffness related to the shaking palsy, proposed to rename the disease in honor of his English colleague.[Tagle, 2019]

In "An essay on the shaking palsy" James Parkinson referred to the shaking palsy as an "Involuntary tremulous motion, with lessened muscular power, in parts not in action and even when supported; with a propensity to bend the trunk forwards, and to pass from a walking to a running pace: the senses and intellects being uninjured"[Parkinson, 2002]. Nowadays the definition for Parkinson's disease establishes it as a progressive multi-system neurodegenerative disorder more common among people in later years of life, and several studies refer to it as the second most common neurodegenerative illness worldwide.[Sveinbjornsdottir, 2016]

In modern medicine it is a common practice among physicians to refer to the mnemonic TRAP (Tremor, Rigidity, Akinesia and Postural Instability) when diagnosing the Parkinson's disease, which includes definitions for all its main clinical features. Even though nowadays Parkinson's disease is far more established and recognized, its correct and early diagnosis still represents a challenge as many of its signs and symptoms are often insidious. Even when referring to the TRAP, many of its symptoms can only be identified through some sort of examination that will normally not be considered unless specifically looking for those. Some of the symptoms of Parkinson's disease that are easier to detect are the following:[Frank et al., 2006]

- The Gait of patients with Parkinson's disease. Which is characterized for the forward flexion of the trunk and a decrease in the length and height of step. This characteristics can be noticed even before the appearance of the postural instability, but can easily be confused with the symptoms of the senile gait related to the aging process.
- The development of a masklike face, also known as hypomimia, which is composed of two main factors. A decrease in the blink frequency of the patient and an

increased dabbling or drooling due to the difficulty of swallowing saliva that patients of Parkinson's disease and other parkinsonian illnesses experiment.

- The development of micrographia, which can be identified by a doctor observing a patient do a written task or even by the sufferer himself.
- The suffering of depression, which is the most common psychiatric feature present in the early stages of the Parkinson's disease.
- The development of dementia which is more of a complication common in the later stages of Parkinson's disease.

When talking about Parkinson's disease we must realize that it's early detection has a mayor impact on the ability to control it's development. Nowadays, newer means of technology allow for newer approaches regarding the diagnosis of Parkinson's disease, and so many later papers focus on the possibility of relying on Artificial Intelligence and Machine Learning to establish patterns based on the parametrization of different data related but not limited to aspects such as the phonic and kinematic disorders that are attributed to this illness.

1.2 Objective

The main objective of this paper is to implement a convolutional neuronal network based system for the early detection of Parkinson's Disease based on the parametrization of audio recordings. In order to do so I will base my research on data taken out from Synapseg.

This objective can be therefore subdivided in the following:

1. Get familiarized with the Python programming language and the Jupiter Notebook environment.
2. Learn the possibilities offered by python's deep learning API keras.
3. Research the state of art concerning both the possible structures for Neuronal Networks that could be relevant for this project, as well as previous researches on this topic.
4. Analyze the raw audio data in order to extract the relevant information to then feed into the Neuronal Networks.
5. Implement, train and compare the results of different Neuronal Network's architectures, in order to provide evidence of their effectiveness regarding the early detection of Parkinson's Disease based on the provided information.
6. Draw conclusions based on the results of the experiments and provide this information so further research can be realized if the results end up being encouraging.

To facilitate this project's success I will also establish the following "Machine Learning Project Checklist" as suggested by Aurélien Géron in his book "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow" [Géron, 2019]:

1. Frame the problem and look at the big picture.
2. Get the data.
3. Explore the data to gain insights.
4. Prepare the data to better expose the underlying data patterns to Machine Learning algorithms.
5. Explore many different models and shortlist the best ones.
6. Fine-tune your models and combine them into a great solution.
7. Present your solution.

1.3 Document Structure

Following this introduction the paper will be structured according to the following sections:

- **State Of Art:**
- **Theoretical Bases:**
- **Data:**
- **Conclusion:**
- **Social Implications:**

1.4 State Of Art

When regarding the application of machine learning techniques to the diagnosis of Parkinson's disease there is a wide range of papers that each present their personal take on the matter.

There are several concerns that need to be addressed when approaching this issue from a technological perspective, the first of which would be to decide on the type of data that is going to be fed into the algorithm.

While some papers focus on the parametrization of the speech, such as "Diagnosis of Parkinson's disease using deep learning and cell phone voice recordings" [Aldana, 2019] others choose to analysis different sources of data that they feel could present interesting patterns. One possibility would be to realize a kinematic analysis of the patient, using data describing motion of upper and lower extremities, as implemented in "Artificial intelligence for assisting diagnostics and assessment of Parkinson's disease—A review" [Belić et al., 2019], while several studies in the platform Synapse investigate the possibility of diagnosing Parkinson's Disease through symptoms related to the loss of recent memory as well as the study of the patients manual abilities.

When choosing the data to be sampled for the experiments it is also relevant to consider the pros and cons of selecting numerous sources of data versus choosing to focus on one specific type. For this paper I have decided to focus on studying voice recordings exclusively, as I find those to have great potential. Nonetheless, I will try to identify any possible relations between the outcomes of the machine learning algorithms application, and other data concerning the state of the patients such as several measurements related to the Unified Parkinson's Disease Rating Scale (UPDRS). The idea of utilizing speech analysis to detect and track the deterioration associated to the evolution of the disease comes from sysmtoms such as Dysarthria that affect this functions. [cnn,]

After choosing the data, it needs to be analyzed in order to separate all the relevant information from the useless bits and noise, for that we will focus on analyzing sets of traits for the data that have previously proven good efficacy when utilizing IA and Deep learning. This will be later discussed on the **Data** section of this document.

The last thing to do would then be to choose a Neuronal Network architecture to implement, but this will be furthered discussed in the following **Theoretical Bases** section.

2 Data

In order to apply any Machine Learning algorithm we need a data set, in this case I will be utilizing one extracted from the Synapse platform. This data set provides 74 different data from a total of 779 participants, all of which has been previously anonymized. From the initial data I will only feed values for the following 40 attributes to the Neuronal Network:

I have selected all the parameters related to the voice recordings as well as the sex and current age which I believe could be relevant for the classification process.

Some of the data from the data set that will not be fed to the Neuronal Network relates to the previously mentioned UPDRS (Unified Parkinson's Disease Rating Scale) and will be utilized later on in order to draw any possible relations between those and the outcomes of the Neuronal network. As this data is not directly related to the voice recordings I have not considered it relevant for the inputs of the Neuronal network which affects the gradient.

Previous to its use in the training and validation of the Neuronal Network the data set must be load and go through several processes. In order to load the data I will use python's Pandas library, which will allow me to generate panda's data frames from the original coma separated CSV files.

2.1 Cleaning The Data Set

The process of cleaning the data will prevent the Neuronal Network from learning any incorrect patterns that may arise of such data and it makes the process of learning more efficient as the incorrect instances would slow down the learning process pushing the weights and bias in wrong directions.

I started the data cleaning process by simply removing all the instances for which the attribute "voice_code", which can take the values "ok" or "bad", had de "bad" value, as those do not include any values for any of the "features" which are to be imputed in the Neuronal Network. To do so I implemented the following function:

```
1 def eliminateBadVoiceCodeElements(data_frame):
2     valid_voice_code_elemnts = data_frame.copy()
3     valid_voice_code_elemnts = valid_voice_code_elemnts.
4     drop(valid_voice_code_elemnts[valid_voice_code_elemnts["
5     voice_code"] == "bad" ].index)
6     return valid_voice_code_elemnts
```

And so before applying the function to the data set we had the following count for the values of the "voice_code" attribute:

```
1 ALL_DATA["voice_code"].value_counts()
2
3
```

```
ok      747
bad      32
Name: voice_code, dtype: int64
```

And after applying the function to the original data set we ended up with a total of 717 instances after deleting the 32 with value "bad" for the attribute "voice_code".

In addition of the previous, in order to further clean the data I will proceed by deleting any instances that have an undefined value for any of the attributes. Although in some scenarios those could be considered a valid value for some attributes, given the nature of our features I rather considered them to be failures in the readings. When eliminating the rows with undefined values I did not just limit myself to erasing the ones with an undefined value in one or more of the attributes that would be imputed on the Neuronal Network but rather I eliminated any row with at least one undefined values in any attribute because when considering this to be a failure in the readings I could not trust that the error did not reflect in the values of the other attributes.

This process of deleting the rows with undefined values can be easily done by invoking the dropna method of the pandas dataframe class.

Then, after further analyzing the data I realized that some of the instances had an empty string value for the gender and as I do believe that this could be a significant attribute to input in the neuronal network i decided such value was not valid within this project's frame and so proceeded with deleting all the instances of data with this value in their sex attribute. In order to do so I implemented the following function:

```
1
2 def eliminateEmptySexElements(data_frame):
3     valid_voice_code_elemnts = data_frame.copy()
4     valid_voice_code_elemnts = valid_voice_code_elemnts.drop(
5         valid_voice_code_elemnts[valid_voice_code_elemnts["
6         voice_code"] == "" ].index)
```

Finally I also deleted any instances for which the current_age attribute was lower than the years_since_first_symptom, as such a condition is not logically possible, and also removed all instances with less than 20 as their current_age value. Which basically eliminated some instances that had a value of 0 for current_age. I did so with the following function:

```
1
2 def checkAgeIsGraterThanSysptoms(data_frame):
3     valid_data_frame = data_frame.copy().reset_index()
4     indexes = []
5     for index in range(len(valid_data_frame)):
6         if valid_data_frame["current_age"][index] <
7         valid_data_frame["years_since_first_symptom"][index] or
8         valid_data_frame["current_age"][index] < 20:
9             indexes.append(index)
10    valid_data_frame.drop(indexes)
11    del valid_data_frame["index"]
12    return valid_data_frame
```

2.2 Normalizing The Data Set And Transforming Non Numerical Values

Normalizing the data previously to training a Neuronal Network with it bares the benefit of speeding up the learning rates by leading to a faster convergence, this is archived by adjusting the range of values of the different attributes to be the same.

In my case I first implemented the min max technique which consists in transforming for every feature it's minimum value to a zero, it's maximum value to a one, and then every other value to a decimal between this two.

$$x \forall Z$$
$$f(x) = (x - MIN(Z)) / (MAX(Z) - MIN(Z))$$

The only significant downside to this technique is the fact that it does not handle outliers well. And according to their plots this would be a problem for several of the features.

We can see the previous problematic represented in the following plot:

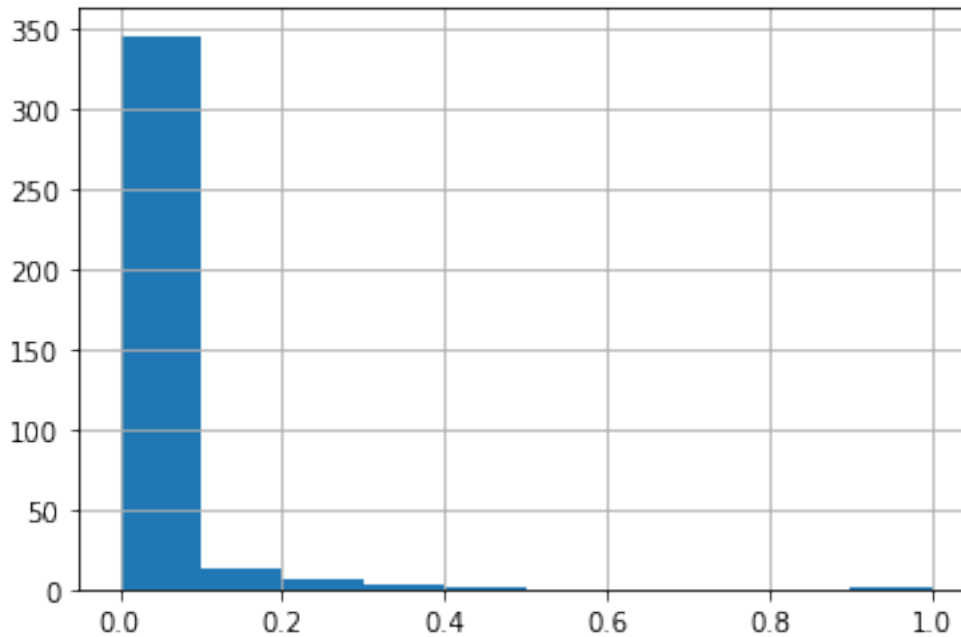


Figure 1: Histogram for a normalized attribute with outliers. Feature02

In order to minimize the impact of applying the min max technique to those attributes with outliers I decided to apply feature clipping which caps a all features above and below the given boundaries to a fixed value, this on time will help spread the feature values over the entire range. Although this is a helpful technique I did not considered it appropriate for the features which plot shape was similar to the one in Figure 1 and only aplyed it to the features with plots that looked like the following figure.

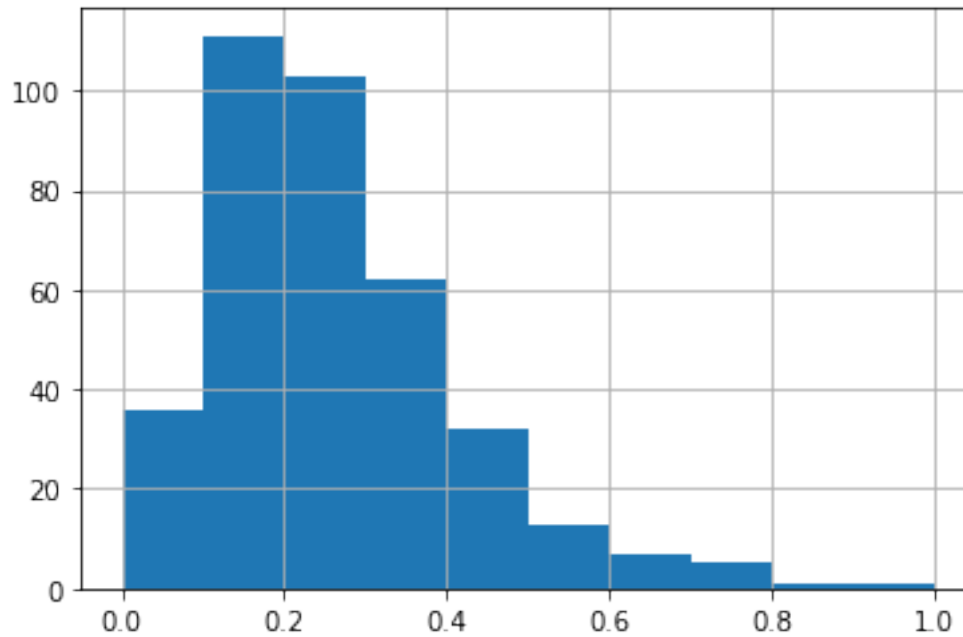


Figure 2: Histogram for a normalized attribute with outliers. Feature24

Which after the feature clipping ended up looking like this:

Figure 3: Histogram for a normalized and clipped attribute. Feature24

Finally to the several attributes with a value distribution similar to the one presented by "feature02" I thought about applying Log Scaling, which aims to compress a wide range of values into a smaller range and is nice to use when a handful of the values repeat many times as happens for those attributes.

$$x \in Z$$

$$f(x) = \log(x)$$

It was after my initial attempt that I realized that many of those attributes have features with zero and negative values which actually complicates the application of the Log Scaling and so I decided to try using the RobustScaler from Python's sklearn library.

At this point in time I also assigned numerical values to "sex", which happens to be the only non numerical attribute that was selected as input for the Neuronal Network. This particular attribute only has two possible values after the cleaning process which are "F" for female participants and "M" for male ones, and so the following function assigns the value 1 for male participants and 0 for female ones.

```

1
2
def normalizeSex(data_frame):

```

```
3     current_dataframe = data_frame.copy().reset_index()
4     n_sex = []
5     for index in range(len(current_dataframe)):
6         if current_dataframe["sex"][index] == "M":
7             n_sex.append(0)
8         else:
9             n_sex.append(1)
10    current_dataframe["sex"] = n_sex
11    del current_dataframe["index"]
12    return current_dataframe
13
```

2.3 Splitting The Data Set Into Training, Test and Validation

2.4 Generating The Labels

2.5 Balancing The Data Set

3 Theoretical Bases

In this paper we will be working within the field of Machine Learning, so in order to describe it I will refer to Tom Mitchell, who referred to Machine Learning as follows:

"A computer program is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E ."

- Tom Mitchell, 1997 [Géron, 2019]

For this paper I will focus on implementing a Convolutional Neuronal Network (CNN)

The difference between ML and Deep Learning,

While machine learning

Applying ML techniques to dig into large amounts of data can help discover patterns that were not immediately apparent. This is called data mining.

RNNs, CNNs, or Transformers for voice recognition

Types of Machine Learning Systems

Whether or not they are trained with human supervision (supervised, unsupervised, semisupervised, and Reinforcement Learning)

Whether or not they can learn incrementally on the fly (online versus batch learning)

Whether they work by simply comparing new data points to known data points, or instead by detecting patterns in the training data and building a predictive model, much like scientists do (instance-based versus model-based learning)

Supervised learning.

In supervised learning, the training set you feed to the algorithm includes the desired solutions, called labels

Batch learning

In batch learning, the system is incapable of learning incrementally: it must be trained using all the available data. This will generally take a lot of time and computing resources, so it is typically done offline. First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned. This is called offline learning.

Instance-based Learning

This is called instance-based learning: the system learns the examples by heart, then generalizes to new cases by using a similarity measure to compare them to the learned

examples (or a subset of them)

Problems to be faced with the data:

Insufficient Quantity of Training Data.

Even for very simple problems you typically need thousands of examples, and for complex problems such as image or speech recognition you may need millions of examples.

Nonrepresentative Training Data

In order to generalize well, it is crucial that your training data be representative of the new cases you want to generalize to. This is true whether you use instance-based learning or model-based learning.

It is crucial to use a training set that is representative of the cases you want to generalize to. This is often harder than it sounds: if the sample is too small, you will have sampling noise (i.e., nonrepresentative data as a result of chance), but even very large samples can be nonrepresentative if the sampling method is flawed. This is called sampling bias.

Poor-Quality Data Synapse tiene un control de calidad, columna que categoriza el audio en correcto o no

Irrelevant Features As the saying goes: garbage in, garbage out. Your system will only be capable of learning if the training data contains enough relevant features and not too many irrelevant ones. A critical part of the success of a Machine Learning project is coming up with a good set of features to train on. This process, called feature engineering, involves the following steps:

Feature selection (selecting the most useful features to train on among existing features)*

Feature extraction (combining existing features to produce a more useful one as we saw earlier, dimensionality reduction algorithms can help)

Creating new features by gathering new data

*Visualization algorithms are also good examples of unsupervised learning

A related task is dimensionality reduction, in which the goal is to simplify the data without losing too much information. One way to do this is to merge several correlated features into one. For example, a car's mileage may be strongly correlated with its age, so the dimensionality reduction algorithm will merge them into one feature that represents the car's wear and tear. This is called feature extraction.

Problems to be faced with the algorithm:

In Machine Learning this is called overfitting: it means that the model performs well on the training data, but it does not generalize well.

Overfitting happens when the model is too complex relative to the amount and noisiness

of the training data. Here are possible solutions:

*Simplify the model by selecting one with fewer parameters (e.g., a linear model rather than a high-degree polynomial model), by reducing the number of attributes in the training data, or by constraining the model. (regularization)

Gather more training data.

Reduce the noise in the training data (e.g., fix data errors and remove outliers).

*The amount of regularization to apply during learning can be controlled by a hyperparameter. A hyperparameter is a parameter of a learning algorithm (not of the model). As such, it is not affected by the learning algorithm itself; it must be set prior to training and remains constant during training. If you set the regularization hyperparameter to a very large value, you will get an almost flat model (a slope close to zero); the learning algorithm will almost certainly not overfit the training data, but it will be less likely to find a good solution. Tuning hyperparameters is an important part of building a Machine Learning system (you will see a detailed example in the next chapter).

As you might guess, underfitting is the opposite of overfitting: it occurs when your model is too simple to learn the underlying structure of the data.

Select a more powerful model, with more parameters. Feed better features to the learning algorithm (feature engineering). Reduce the constraints on the model (e.g., reduce the regularization hyperparameter).

Training:

A better option is to split your data into two sets: the training set and the test set. As these names imply, you train your model using the training set, and you test it using the test set. The error rate on new cases is called the generalization error (or out-of-sample error), and by evaluating your model on the test set, you get an estimate of this error. This value tells you how well your model will perform on instances it has never seen before.

A common solution to this problem is called holdout validation: you simply hold out part of the training set to evaluate several candidate models and select the best one. The new held-out set is called the validation set (or sometimes the development set, or dev set).

Esto es más para saber cómo hacerlo que para incluirlo en la memoria.

More specifically, you train multiple models with various hyperparameters on the reduced training set (i.e., the full training set minus the validation set), and you select the model that performs best on the validation set. After this holdout validation process, you train the best model on the full training set (including the validation set), and this gives you the final model. Lastly, you evaluate this final model on the test set to get an estimate of the generalization error.

cross-validation??

<https://towardsdatascience.com/step-by-step-vgg16-implementation-in-keras-for-beginners-a833c686ae6c>

4 Conclusion

4.1 Social Implications

5 Possible Future Work Lines

Yet another important unsupervised task is anomaly detection—for example, detecting unusual credit card transactions to prevent fraud, catching manufacturing defects, or automatically removing outliers from a dataset before feeding it to another learning algorithm. The system is shown mostly normal instances during training, so it learns to recognize them; then, when it sees a new instance, it can tell whether it looks like a normal one or whether it is likely an anomaly

References

- [cnn,] Identification of parkinson’s disease from speech using cnns and formant measures.
- [Aldana, 2019] Aldana, B. G.-B. (2018-2019). *Diagnóstico de la enfermedad de Parkinson usando deep learning y grabaciones de voz mediante teléfono móvil*. Universidad Politécnica de Madrid.
- [Belić et al., 2019] Belić, M., Bobić, V., Badža, M., Šolaja, N., Đurić-Jovičić, M., and Kostić, V. S. (2019). Artificial intelligence for assisting diagnostics and assessment of parkinson’s disease—a review. *Clinical neurology and neurosurgery*, 184:105442.
- [Frank et al., 2006] Frank, C., Pari, G., and Rossiter, J. P. (2006). Approach to diagnosis of parkinson disease. *Canadian family physician*, 52(7):862–868.
- [Géron, 2019] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O’Reilly Media, Inc.
- [Parkinson, 2002] Parkinson, J. (2002). An essay on the shaking palsy. *The Journal of neuropsychiatry and clinical neurosciences*, 14(2):223–236.
- [Sveinbjornsdottir, 2016] Sveinbjornsdottir, S. (2016). The clinical symptoms of parkinson’s disease. *Journal of neurochemistry*, 139:318–324.
- [Tagle, 2019] Tagle, D. P. (2019). Historia de la enfermedad de parkinson.