Exploring the feasibility of DNN models for the quantitative discrimination between different conformational species of α-synuclein

Camillo Nicolò De Sabbata, Gianluca Radi, Thomas Berkane
Department of Computer Science, EPFL, Switzerland

Abstract—**Determining the quantities of different conformational species of α-synuclein in body fluids can be of use for the early diagnosis of Parkinson's Disease in humans. Machine Learning techniques, trained on data acquired through infrared spectroscopy, are thus an interesting prospect to predict these quantities. This project uses supervised methods to solve two tasks: the classification of spectral signatures of different conformational species of α-synuclein and a regression task for quantitative differentiation of distinct conformational species when present in different ratios. The dataset consists of absorption signatures obtained by infrared spectroscopy on PBS buffer containing different conformational species in different ratios. For the first task, we trained an MLP model on spectra containing a unique protein species and achieved an accuracy on the test set of about 95%. For the second task, we trained both an MLP and an Extremely Randomized Trees model on spectra containing different ratios of protein species; the former performing better with a Mean Squared Error of 0.014 and a Mean Absolute Error of 0.061**

I. Introduction

Parkinson's Disease (PD) is the fastest-growing neurodegenerative disorder, with almost 6.2 million cases worldwide[1]. The condition of PD is concomitant with many clinical motor and non-motor symptoms, eventually necessitating assisted care for PD patients with the disease progression[2]. Despite the devastating consequences and impact of PD, there is neither a cure that can halt the progression of the disease nor any diagnostic method for the early detection yet to tackle this issue. The standard diagnosis for Parkinson's is clinical. But this happens at an advanced stage of the disease, which thwarts any effort to treat or delay the progression of the disease. The knowledge of the disease pathology and the understanding of the early events of the disease can help us develop new diagnostic tools to assist in the early preclinical detection of the disease.

One important hallmark of PD is the presence of Lewy Bodies (LBs) which are intraneuronal inclusions in different parts of the brain[3]. These inclusions are enriched with α-syn aggregates, a 140-aa, 15 kDa protein coded by SNCA gene, which is predominantly found in the brain regions-neocortex, hippocampus, substantia nigra, thalamus, and cerebellum[4]. Over the years, the processes of α-syn aggregation, propagation, and transmission of such misfolded α-syn between the neuron cells and its role in PD pathogenesis have been well understood. The level of α-syn presence in the brain depends on the balance between its synthesis, aggregation, and removal rate[5]. A disruption in this balance could lead to the accumulation of the protein in abnormal levels that could favor the formation of misfolded oligomeric or fibrillar aggregates[6].

Since the α-syn misfolding is one of the early events in the pathology and progression of PD, it could be utilized as an early structural biomarker for diagnosing PD in humans at the preclinical stage itself. Such techniques target to sense the conformational changes of proteins from their healthy random or α-helix monomeric state to the pathological cross beta-sheet enriched species. The discovery of α-syn in the body fluids like CSF and blood made this research direction further interesting as it could make the diagnostic method faster, more accessible, and less invasive[7].

From the absorption signatures of α-syn monomers, oligomers, and PFFs, obtained by infrared spectroscopy, differences in secondary structure can be analyzed using methods like second derivative analysis, Fourier deconvolution, and curve fitting, as shown in Fig.1. The cross β-sheet structures have predominant absorbances in the lower wavenumbers between 1620 and 1640 cm-1, whereas the random structures or α-helix absorb between 1640-1670 cm-1. But the classical spectral analysis falls short in resolving highly similar yet fundamentally different spectral signatures of other conformational species of the same protein when present simultaneously on the sensor surface. This project aims to solve this challenge by augmenting the sensor with a Neural Network model to deliver accurate and quantitative differentiation of distinct conformational species in different ratios.

II. Data Analysis

A. Data Obtainment And Selection

The data was provided to us by Deepthy Kavungal, who supervised this project. It consists of CSV files whose values represent absorption spectra: each row corresponds to a wave number, and each column corresponds to a different timepoint. Thus in these files, each column corresponds to one data point. Wave numbers go from 1450 cm^{-1} to 1700 cm^{-1}, with a step size of 2 cm^{-1}, and there are about 1000 timepoints in total.

Different CSV files correspond to spectra with varying proportions of Oligomers and PFFs, with possibilities being: $0\%-100\%$, $25\%-75\%$, $40\%-60\%$, $50\%-50\%$, $60\%-40\%$,

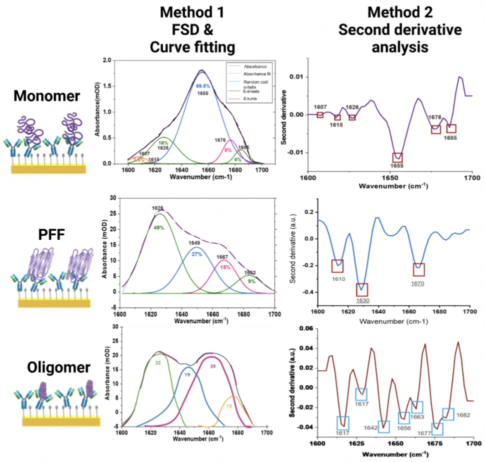

Fig. 1. Secondary structure analysis of α-syn monomers, oligomers and PFFs using Fourier deconvolution and curve fitting, and second derivative analysis

$75\% - 25\%$, $100\% - 0\%$. Note that these were the only concentrations mixes of Oligomers and PFFs whose data was available, as obtaining spectra for a new proportion is a lengthy and costly process.

We also had access to spectra containing Monomers, which could be used as a backup plan as they are easier to distinguish from PFFs, but we did not end up using this data as our results were sufficient.

Moreover, we had access to time-averaged (5x or 10x) versions of the same data, which reduces noise but also reduces our dataset's size. In the end, we decided to use the non-time-averaged data as it gave the best performance.

B. *Train/test set splitting and cross-validation*

Two ways to split the data in train and test sets were utilized. The first way consists of using 80% of the spectra for training and 20% of the spectra for testing, with all the different Oligomes and PFFs proportions present in both sets. The second way involves training on all the spectra for all the proportions except for some that are used for testing. This split aims to test if our model generalizes well to unseen proportions (which corresponds to the actual case usage).

In both cases, cross-validation was then applied to select the best model and hyperparameters.

III. Models

A. *Classification task*

The first task we addressed was to discriminate between the spectral signatures of two conformational species of α-synuclein, namely PFFs and Oligomers. To achieve this, we built a simple Multilayer Perceptron (MLP) with three layers of 10 nodes each, logistic activation function, adamax optimization algorithm, and MSE loss function, which achieved an accuracy of around 95% in 4-fold cross-validation. Note that we decided to implement the MLP without using already available models from libraries, as we wanted to have the highest degree of control regarding the architecture of the MLP itself. However, a simple architecture proved to be sufficiently good, as already shown.

B. *Regression task*

We then proceeded to assess the feasibility of predicting the concentration mixes of Oligomers and PFFs contained in the PBS buffer, starting from their absorption spectra. To do that, we decided to *simulate* the real case usage of the desired model. That is, having a model trained on a specific dataset, which should be able to predict a concentration of PFFs and Oligomers that was not necessarily contained in the training set. For this reason, our goal for this task was to find a model with reasonable performance, which would be trained on an increasingly larger datasets of concentrations of Oligomers and PFFs and then tested on specific concentrations of Oligomers and PFFs not contained in the training set.

C. *Multilayer Perceptron for Regression*

We opted for an MLP for the regression task as well, as it seemed somewhat similar to the classification one. Indeed, the presented classification task can be interpreted as a regression problem, where the model is expected to output a 0 for mixtures of pure Oligomers and a 1 for mixtures of pure PFFs. Hence, since the MLP proved to be a solid model for that type of problem, we employed it also for the regression. Note, however, that the MLP used for the regression part has been imported from the sklearn library. This MLP has 4 hidden layers composed of 10 nodes each, with logistic activation function and adam optimization algorithm. We then proceeded with the training and testing of the model.

We started by creating 2 datasets, the first composed of 350 datapoints for concentration $40\% - 60\%$ and the second composed of 350 datapoints for concentration $60\% - 40\%$ and we concatenated them. We then bootstrapped 100 subsamples of 200 datapoints each from the aforementioned dataset. These subsamples will be used for testing the model, wheras for training the following subsets of concentrations will be employed:

-*S1* = $\{0\% - 100\%, 100\% - 0\%\}$

-*S2* = $\{0\% - 100\%, 50\% - 50\%, 100\% - 0\%\}$

-*S3* = $\{0\% - 100\%, 25\% - 75\%, 50\% - 50\%, 75\% - 25\%, 100\% - 0\%\}$

With such subsets, the model will not train on the data representing the $40\% - 60\%$ and $60\% - 40\%$ concentrations, hence simulating the real case scenario already mentioned.

Firstly, we trained the model on *S1*, tested it on the 100 datasets, computed the Mean Squared Error (MSE), Mean Absolute Error (MAE), and Maximum error for each of the 100 datasets, and yielded the mean of the mentioned errors. Not surprisingly, the predictions were not satisfactory. Indeed, the MSE was around 0.19, the MAE was around 0.42, and the Maximum Error was around 0.58.

Secondly, we trained the model on *S2*, tested it on the 100 datasets, computed the MSE, MAE, and MAX errors for each of the 100 datasets, and yielded the mean of the mentioned errors again. At this point, a better result was obtained, as the

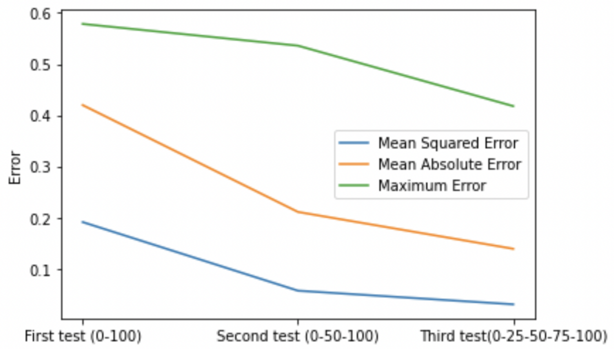

Fig. 2. Improvement of errors when adding new concentrations data

MSE was around 0.06, the MAE was around 0.21, and the MAX error was around 0.54.

Finally, we trained the model on *S3* and repeated the process already discussed. As expected, we got the best results, as the MSE was around 0.03, the MAE was around 0.14, and the MAX Error was around 0.42.

D. Parameter Tuning

Once we proved the feasibility of the procedure, we proceeded by tuning the parameters of the Neural Network. We tuned the number of layers going from 2 to 20, the number of nodes in each layer going from 2 to 20, and the step size of the optimization from 10^{-3} to 10^{-5}. The best result, which corresponds to the Negative Mean Absolute Error of 0.061 and a Mean Squared Error of 0.014 in 5-fold cross-validation, was obtained with an MLP of depth 6, width 36, and regularization term (α) 10^{-5}.

E. Extremely Randomized Trees

As a point of comparison with the Neural Network approach, we tried another method for the regression task, using the Extremely Randomized Trees (ERT) model[8], a variation of the Random Forest model which seeks to avoid overfitting.

Indeed, a concern we kept in mind while selecting a model to use was that we wanted it to generalize well to unseen proportions of Oligomers and PFFs. As the ERT model does just this and has been used successfully for a similar application[9], we gave it a try.

The input to the model consisted of the spectra. The number of estimators, maximum depth, minimum samples split, the maximum number of features, and minimum samples leaf were selected using 5-fold cross-validation for training the model. Optuna[10] was used to select these parameters.

A Mean Absolute Error of about 0.27 was achieved on the test set. Thus, as Table 1 shows, this model does not perform as well as the Neural Network.

IV. Results

As previously stated, the main goal of this project was to assess whether machine learning techniques could be employed to carry out the quantitative discrimination between conformational species of α-synuclein. Therefore, that was the central aspect on which we focused our efforts. For this reason, we concentrated on trying to prove, although empirically, that a model capable of yielding good results was possible to build. We also tried to illustrate how such a model could be trained. That is, we aimed at understanding which type of data can be helpful for training the model. However, we did not strive to perfect our model to achieve exceptional performance, which was not our primary interest.

TABLE I
Comparison of ERT and MLP results

Model	MAE
ERT	27%
MLP	6.1%

Regarding the project's feasibility, we have shown that a simple MLP can discriminate between the different spectral signatures of Oligomers and PFFs. Not only that, but we were also able to produce a model (yet another MLP) capable of performing regression to predict the concentrations mixes of α-synuclein conformational species. In this regard, some analysis regarding the behavior of the model with respect to the training data was carried out.

The first outcome of such analysis is that the more datasets of different concentrations mixes are used for training, the higher the model's performance. Nonetheless, it did not seem helpful to have many data points for each concentrations mix, nor to use the time-averaged data. That is because the model employed is a neural network, hence standard pre-processing techniques are not required and raw data can be directly used for training. Note that this also allows to reduce the time required for the extraction of the data.

Another noteworthy aspect is that the distribution of the concentrations employed for training makes a difference. In particular, if the MLP is trained with the absorption signatures of concentrations that are *equidistant* (such as $0\% - 100\%$, $50\% - 50\%$, $100\% - 0\%$), it is capable of generalizing better and tend not to overfit when presented with a new concentration it has not been trained on. This is why we ended up choosing the presented subsets for the simulation of the real case usage of the model.

Lastly, another particular behavior we observed is that when the test concentrations are *close* to the concentrations used for the training, the MLP achieves higher performances. This explains why the gain of the MLP trained on *S2* on the one trained on *S1* is higher than the gain of the MLP trained on *S3* on the one trained on *S2*. Indeed, $50\% - 50\%$ is closer to $40\% - 60\%$ and $60\% - 40\%$ than $25\% - 75\%$ and $75\% - 25\%$ are. So, adding the last two concentrations to the training set is not as effective as adding the first one when testing on $40\% - 60\%$ and $60\% - 40\%$.

V. Conclusions

In this project, we explored the feasibility of applying machine learning techniques for the quantitative discrimination between different conformational species of α-synuclein.

We firstly focused on the classification of spectral signatures of two conformational species: Oligomers and PFFs. We achieved this with a Multilayer Perceptron (MLP) with 4 layers composed of 10 nodes each, yielding an accuracy of around 95%.

Then, we addressed a regression problem, namely, predicting the concentrations mixes of Oligomers and PFFs from their absorption signatures. The model employed for such a task was an MLP, different from the first one, with a slightly more complex architecture. With the available data, we obtained a Negative Mean Absolute Error of 0.061 in 5-fold cross-validation.

Additionally, we built an Extremely Randomized Trees model for the regression task to investigate whether classic machine learning techniques would have yielded similar results to Neural Networks. However, our findings showed that the MLP performed better, with a Mean Absolute Error more than 4 times smaller.

Finally, we tried to understand how to obtain better results from the MLP for the regression problem. The findings already discussed in the previous section seem to suggest that higher performance is achieved when training the model on data representing concentrations mixes that are *close* and *equidistant* within each other. Hence, training the model on the data for concentrations mixes $0\% - 100\%$, $10\% - 90\%$, $20\% - 80\%$, $30\% - 70\%$, $40\% - 60\%$, $50\% - 50\%$, $60\% - 40\%$, $70\% - 30\%$, $80\% - 20\%$, $90\% - 10\%$, and $100\% - 0\%$ can possibly result in an increase of performance.

References

[1] E. R. Dorsey et al., “Global, regional, and national burden of parkinson’s disease, 1990–2016: A systematic analysis for the global burden of disease study 2016,” *The Lancet Neurology*, vol. 17, pp. 939–953, 2018.

[2] G. Ayano, “Parkinson’s disease: A concise overview of etiology, epidemiology, diagnosis, comorbidity and management,” *Journal of Neurological Disorders*, vol. 4, 2016.

[3] D. J. Moore, A. B. West, V. L. Dawson, and T. M. Dawson, “Molecular pathophysiology of parkinson’s disease,” *Annual Review of Neuroscience*, vol. 28, pp. 57–87, 2005.

[4] F. N. Emamzadeh, “Alpha-synuclein structure, functions, and interactions,” *J Res Med Sci*, vol. 21, 2016.

[5] C. L. Kragh, K. Ubhi, T. Wyss-Corey, and E. Masliah, “Autophagy in dementias,” *Brain Pathology*, vol. 22, pp. 99–109, 2012.

[6] H. A. Lashuel, C. R. Overk, A. Oueslati, and E. Masliah, “The many faces of -synuclein: From structure and toxicity to therapeutic target,” *Nature Reviews Neuroscience*, vol. 14, pp. 38–48, 2013.

[7] O. M. A. El-Agnaf et al., “-synuclein implicated in parkinson’s disease is present in extracellular biological fluids, including human plasma,” *The FASEB Journal*, vol. 17, pp. 1–16, 2003.

[8] P. Geurts, D. Ernst, and L. Wehenkel, “In-situ ultra-sensitive infrared absorption spectroscopy of biomolecule interactions in real time with plasmonic nanoantennas,” *Machine Learning*, vol. 63, pp. 3–42, 2006.

[9] Y. Uehara, S. Ueno, and H. Amano-Takeshigeet al., “Non-invasive diagnostic tool for parkinson’s disease by sebum rna profile with machine learning,” *Scientific Reports*, vol. 11, 2021.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, *Optuna: A next-generation hyperparameter optimization framework*, 2019. arXiv: 1907.10902 [cs.LG].