*Bioinformatics*, YYYY, 0–0

doi: 10.1093/bioinformatics/xxxxx

Advance Access Publication Date: DD Month YYYY

Manuscript Category

|  |
| --- |
| Python/Flask  Web application for the Systemic Lupus Erythematosus (SLE) project using Flask  Theodor Rumetshofer1,\*  1Department of Biology, Box 118, 22100 Lund, Sweden  \*To whom correspondence should be addressed.  Associate Editor: XXXXXXX  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Abstract  **Summary:** Web frameworks using Flask provide a powerful tool for data handling and analysis. This application was applied on a test data set of the Systemic Lupus Erythematosus (SLE) project at Lund University which allows all collaboration access to an easy-to-use analysis tool for PCA and hierarchical clustering on cognitive test data. This provides a quick and easy evaluation and interpretation of the data in standard web browsers also for non-advanced users and collaborators of the project.  **Availability:** This program is written in Python using Flask and it is available in GitHub (<https://github.com/TheoRum/BINP29_project_database>). This software is not freely available only for non-commercial users.  **Contact:** theodor.rumetshofer@gmail.com  **Supplementary information:** Supplementary data are available at GitHub (<https://github.com/TheoRum/BINP29_project_database>) |

# Introduction

Systemic Lupus Erythematosus (SLE) is a rare chronic autoimmune disease that predominantly affects healthy women of child-bearing age1. It affects multiple organ systems including the central nervous system which results in neuropsychiatric (NP) symptoms, e.g. cognitive dysfunction, mood disorder and cerebrovascular diseases2. Due to the heterogeneous clinical symptoms in SLE, differential diagnostic and interdisciplinary research is necessary.

In 2013, Lund University launched an interdisciplinary research project on SLE using magnetic resonance imaging (MRI), neuropsychological testing, clinical and laboratory testing. Over the years, various types of data were collected and stored in different institutes and formats.

The goal of this study it to create a central storage of the data to allow collaborators access to the different types of information. Further, to filter and analysis data in a user-friendly graphical user interface (GUI). For this reason, a web framework with flask was implemented. The following methods section will outline the technical background of the applied software packages followed by the implemented features and results. The last section will summarize the project, shows limitations and give an outlook over the next steps.

# Methods

Due to the requirements in the last chapter, a web application framework was implemented. For this purpose, *Flask (1.1.1)*, *Python (3.8.1)*, was used, which allows an easy integration of background application to a web browser. This includes the following packages for handling templates *jinja2 (2.11.1),* input forms *flask-wtf (0.14.3)* and a toolkit for the web server gateway interface *Werkzeug (1.0.0).* These tools together allow processing of the data which is stored in a csv-file in the backend. As a frontend browser Google Chrome (80.0.3987.132) was used.

For the analysis part, principal component analysis (PCA) and hierarchical clustering were implemented using *scikit-learn (0.22.2)* and *seaborn (0.10.0)*, respectively. This admits a quick and easy evaluation and interpretation within the web browser. Further, these two methods seem to be very robust against the heterogeneous datasets.

For this analysis a manually curated test dataset was used from the SLE project which includes patient information (e.g. age, clinical label, quality of life score) and cognitive test data using the CNS vital signs test (CNS-VS)3. Due to general data protection regulations (GDPR), the test data set was fully anonymized.

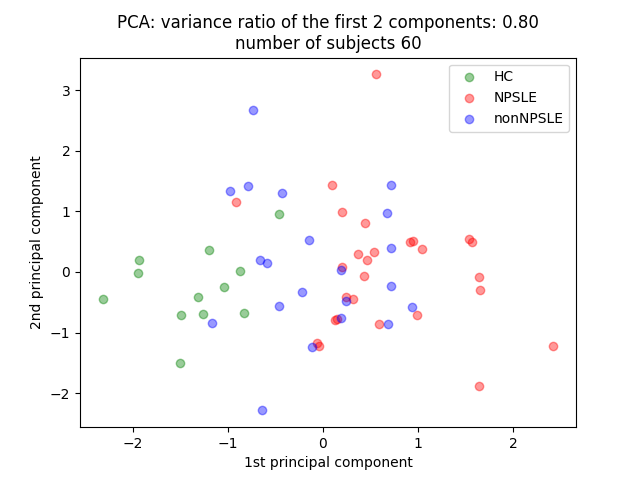
# Features

Flask was used to provide the web framework and access to the dataset which allows a user-friendly access using standard web browsers.

## PCA

The CNS-VS standard scores are a helpful tool to distinguish between healthy controls (HC) and different SLE subgroups with (NPSLE) and without (nonNPSLE) events.

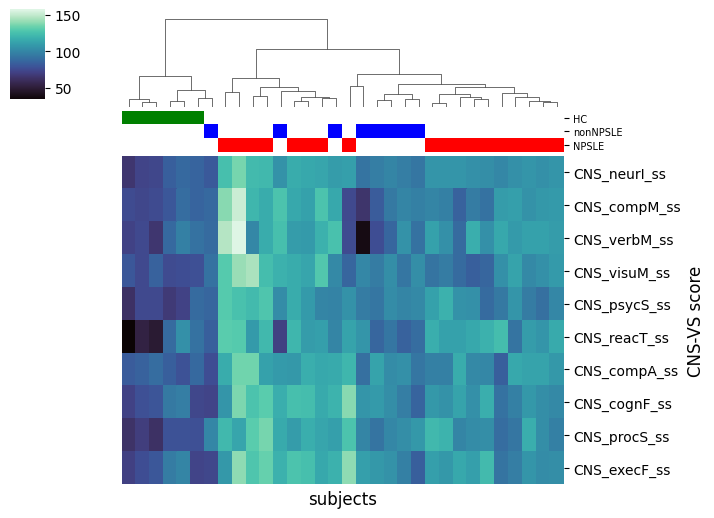
Figure 1 shows the first two PCA components of a filtered test dataset with includes 60 subjects without outliers. Due to the definition of NPSLE subjects, which show overall lower cognitive scores than nonNPSLE and HC, it indicates a larger distance in the components between the other groups. NonNPSLE subjects, which show lower scores than HC but higher than NPSLE, are between both groups.



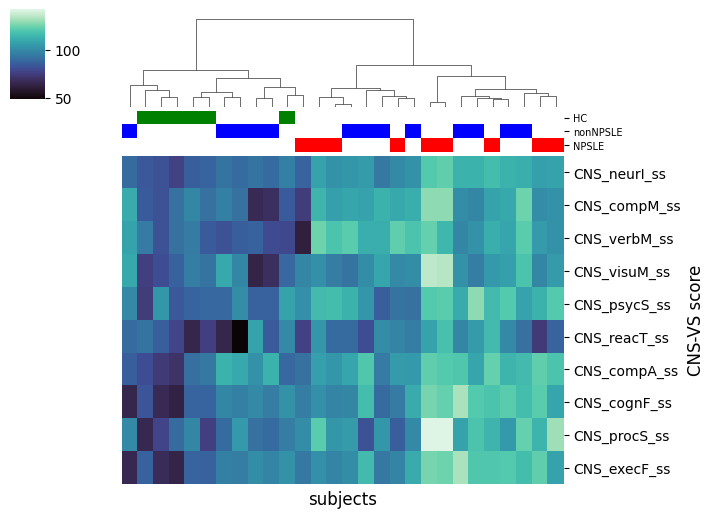
**Fig. 1. PCA of the CNS-VS standard scores** of the filtered test dataset including 60 subjects. The first two components explain 80% of the variance. The three different colors indicate the different clinical labels; HC=healthy controls, NPSLE=neuropsychiatric SLE, nonNPSLE=non-neuropsychiatric SLE

## Clustering

For the hierarchical clustering the CNS-VS standard scores were also used. Figure 2 and 3 show the clustermap of subjects in the test data set with age <40 and >=40 years, respectively. Cluster analysis of subjects below an age of 40 shows homogenous clusters compared to those above or equal to 40 years. We assume that younger HC have slightly higher cognitive values compared to older HC.



**Fig. 2. Hierarchical Ward clustering of subjects <40 years** on the CNS-VS standard scores (y-axis)**.** The clinical labels are shown in the colored bars below the dendrogram; HC=healthy controls, NPSLE=neuropsychiatric SLE, nonNPSLE=non-neuropsychiatric SLE



**Fig. 3. Hierarchical Ward clustering of subjects >=40 years** on the CNS-VS standard scores (y-axis)**.** The clinical labels are shown in the colored bars below the dendrogram; HC=healthy controls, NPSLE=neuropsychiatric SLE, nonNPSLE=non-neuropsychiatric SLE

## Additional features

Further, to allow only project collaborators access to the data a user login was also implemented. As mentioned above only simple queries were implemented in this first prototype which includes outliers, subject age and quality-of-life score. Applying such filters, the web application shows the number of subjects due to the queries and provides a download option in csv-format. However, all filters are tested for invalid input.

# Discussion

This web application with *Flask* including PCA and hierarchical cluster analysis provides a simple and easy-to-use interface also for non-advanced users. Compared to a standard *Python* software, this application does not require any deeper understanding and programming skills.

However, the user interface in very rudimentary and needs a better graphical interface. Also, regarding help and information about wrong input and queries for the users. Further limitations are that just a small amount of the SLE project data was used and instead of a relational database a simple csv-file stored all the information.

To summarize, although *Flask* provides an easy implementation into a *Python* environment the different packages requires a lot of knowledge e.g. in *HTML* and *jinja2*. Nevertheless, this web application is the first step for a central storage and analyses framework in the SLE project at Lund University.

Acknowledgements

Special thanks to all my colleagues of the master’s programme in Bioinformatic at Lund University.

*Conflict of Interest:* none declared.

References

[1] Renau, A., & Isenberg, D., (2012) Male versus female lupus: a comparison of ethnicity, clinical features, serology and outcome over a 30 years period. *Lupus*, 21(10), 1041–1048.

[2] Liang, M. H., Corzillius, M., Bae, S. C., Lew, R. A., Fortin, P. R., Gordon, C., … Winer, J. B. (1999). The American College of Rheumatology nomenclature and case definitions for neuropsychiatric lupus syndromes. *Arthritis & Rheumatism*, *42*(4), 599–608.

[3] Gualtieri CT, Johnson LG. Reliability and validity of a computerized neurocognitive test battery, CNS Vital Signs. *Arch Clin Neuropsychol*. 2006;21(7):623–643.