An analysis of protein patterns present in the saliva of diabetic patients using pairwise relationship and hierarchical clustering*

Airton Soares[1 [0000−0003−2151−5444]], Eduardo Esteves[2,3 [0000−0001−5458−4978]],, Nuno Rosa[2 [0000−0003−4604−0780]], Ana Cristina Esteves[2 [0000−0003−2239−2976]], Anthony Lins[4 [0000−0002−7153−841X]], and Carmelo J. A. Bastos-Filho[1 [0000−0002−0924−5341]]

[1] Universidade de Pernambuco, Recife PE, Brazil
assj@ecomp.poli.br, carmelofilho@poli.br
[2] Universidade Católica Portuguesa, Faculty of Dental Medicine, Center for Interdisciplinary Research in Health (CIIS), Viseu, Portugal
nunorosa33@gmail.com
[3] Universidade da Beira Interior, Faculdade de Ciências da Saúde (UBI), Covilhã, Portugal
https://ubi.pt
[4] Universidade Católica de Pernambuco
anthony.lins@unicap.br

Abstract. Molecular diagnosis is based on the quantification of RNA, proteins, or metabolites whose concentration can be correlated to clinical situations. Usually, these molecules are not suitable for early diagnosis or to follow clinical evolution. Large-scale diagnosis using these types of molecules depends on cheap and preferably noninvasive strategies for screening. Saliva has been studied as a noninvasive, easily obtainable diagnosis fluid, and the presence of serum proteins in it enhances its use as a systemic health status monitoring tool. With a recently described automated capillary electrophoresis-based strategy that allows us to obtain a salivary total protein profile, it is possible to quantify and analyze patterns that may indicate disease presence or absence. The data of 19 persons with diabetes and 58 healthy donors obtained by capillary electrophoresis were transformed, treated, and grouped so that the structured values could be used to study individuals' health state. After Pairwise Relationships and Hierarchical Clustering analysis were observed that amplitudes of protein peaks present in the saliva of these individuals could be used as differentiating parameters between healthy and unhealthy people. It indicates that these characteristics can serve as input for a future computational intelligence algorithm that will aid in the stratification of individuals that manifest changes in salivary proteins.

Keywords: Data mining · Clustering Algorithms · Saliva · Capillary Electrophoresis · Desease Diagnosis

* Supported by Universidade de Pernambuco (UPE) and Universidade Católica Portuguesa (UCP).

1 Introduction

Molecular diagnosis is based on the quantification of RNA [3], proteins [17], or metabolites, whose concentration can be correlated to clinical situations. Usually, these molecules alone are not suitable for early diagnosis or to follow the clinical evolution. Therefore, strategies to evaluate the complete molecular scenario – early diagnosis, diagnosis, and clinical evolution – are necessary.

The potential of proteins for a large-scale diagnosis depends on cheap and preferably non-invasive strategies for screening. A good approach involves bioinformatics strategies and solutions to work with different types of data, from biological-related data to personal and clinical information. Data integration is an asset to predict the pathological status before clinical outcomes.

In the last decade, saliva has been studied as a non-invasive, easily obtainable diagnosis fluid [14]. It is composed of the secretions of the three largest salivary glands (parotid, submandibular and sublingual), smaller salivary glands, crevicular fluid, and contains serum components, transported by blood capillaries, and subsequently transferred by diffusion, transport and/or ultrafiltration. The presence of serum proteins in saliva enhances its use as a systemic health status monitoring tool [2, 10, 11, 19].

Data on salivary proteins associated with disease or health status is already extensive. Our group has studied salivary proteins and produced the SalivaTecDB database (http://salivatec.viseu.ucp.pt/salivatec-db) [1, 15], which is relevant for the identification of proteins that may potentially be associated with specific signatures. SalivaTecDB has currently stored more than 3,500 human salivary proteins.

We recently described an automated capillary electrophoresis-based strategy that allows one to obtain a salivary protein profile – the SalivaPrint Toolkit [4, 7]. Since proteins are separated according to their molecular mass, changes in peak morphology or fluorescence intensity (translated by changes in peak height) correspond to fluctuations in the proteins' concentration or the type of proteins being expressed. The association of saliva protein signatures to different health/disease situations allows us to build a cheap and robust framework for the development of a monitoring tool.

The use of machine learning algorithms on risk disease prediction is already a reality [8, 12, 13, 18]. Clinical data patients integrated with laboratory results can contribute to health/disease monitoring, building the foundations for the development of a risk assessment tool for diagnosis.

In this article, we propose a methodology for the analysis of salivary protein patterns that reflects patients' health status. Using a database of healthy and diabetic individuals protein patterns, we analyzed the association of protein peaks with these patients' health status. The goal is to understand the relationships between the protein profiles and the individual's state of health. This analysis can influence the choice of the learning method for the recognition process.

The remainder of the paper is organized as follows. In Section 2, we explain the methodology deployed in the project, describing the data acquisition process, graphical representation of the obtained data, pre-analysis, the peaks detection

process, the pairwise relationship, and the hierarchical clustering. In Section 3, we have the analysis of the results obtained using the described methodology. In Section 4, we present proposals for future works in the scope of the project. In Section 5, we give our conclusions.

2 Methodology

Aiming at the future development of a computational intelligence algorithm that can differentiate healthy individuals from unhealthy ones, we performed several transformations and analysis procedures on data salivary protein patterns.

2.1 Data acquisition and description

All data used in this study were acquired through capillary electrophoresis, using the Experion™Automated Electrophoresis System (BioRad®) in standard protein chips (Experion™Pro260 Analysis Kit[5]). Total protein concentration was normalized to 500µg/mL, and samples were analyzed in duplicate.

Protein profiles were obtained using the Experion™Software, version 3.20, and exported as an XML file. Once the output file was generated, a Python script extracts the data in the file and generates a CSV (Coma Separated Values) file with 399 signals for each sample. These signals correspond to the fluorescence on each molecular weight measured on the capillary electrophoresis system: 10kDa to 121kDa.

The resulting data set was obtained from 77 individuals, 58 are healthy, and 19 have diabetes. Table 1 shows some examples of the data set structure. The first column contains the identifiers. The last column includes the individual's health status, and the columns between them represent protein weights in kDa, ranging from 10 to 121. The values for the columns are the fluorescence returned from the Experion™Software.

Table 1: Examples of data set entries. The first column contains the identifiers, the last column contains the individual's health status and the columns between them represent protein molecular weights

Sample ID	10.0	10.1	10.2	...	120.3	120.8	121.0	Health status
d1122	-25.914	-23.452	-21.871	...	-1.056	-1.204	-1.332	Healthy
d1127	30.009	25.470	21.091	...	0.756	0.260	-0.472	Healthy
d1132	1.189	0.311	-0.650	...	-0.341	-0.600	-1.043	Healthy
d52	-11.405	-12.425	-13.341	...	-1.025	-1.096	-1.096	Diabetes
d56	-6.595	-6.886	-7.056	...	0.760	0.950	0.950	Diabetes
d59	38.839	25.473	12.131	...	-12.617	-12.682	-12.682	Diabetes

[5] http://www.bio-rad.com/webroot/web/pdf/elsr/literature/10000975C.pdf

2.2 Data visualization

Figure 1 shows the protein profiles for healthy and unhealthy individuals. Fluorescence values have a small variation for healthy individuals than for unhealthy individuals. Also, unhealthy individuals present higher values on the entire scale and a more significant change than healthy individuals. Both groups show the peaks in similar molecular weights. This behavior indicates that the peak height values identified in the signature of an individual's proteins may be used to characterize his health status.

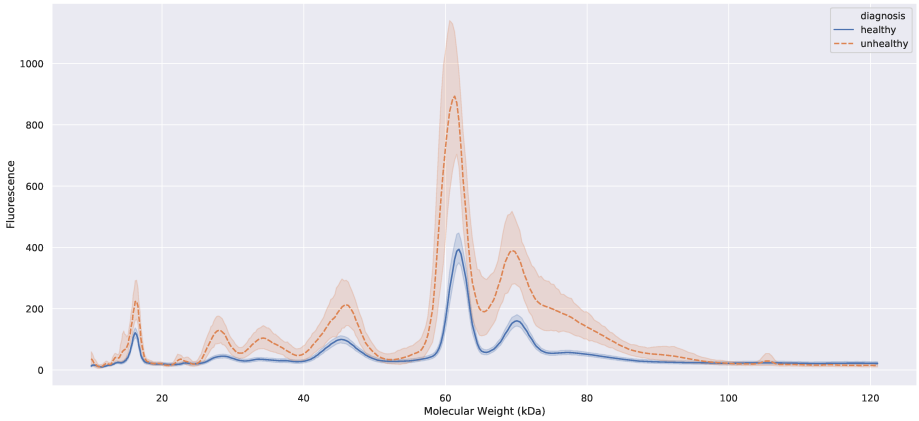


Fig. 1: Representation of the distribution of fluorescence values for each molecular weight (kDa). The average values of the molecular weight's fluorescence for the healthy individuals (n=58) are represented by the blue line and for the unhealthy individuals (n=19) represented by the dashed orange line. The shaded areas around the lines represent the standard deviation of the fluorescence values.

2.3 Data preparation

As can be seen in Table 1, some fluorescence values are negative. Therefore, each row fluorescence values were normalized by adding the absolute value of the smallest value on each row.

The total number of points in each electropherogram is 395. Since the goal of this analysis is to generate valuable information for the selection and calibration of a future machine learning algorithm, the number of features should be as lower as possible without losing significant information [6]. Therefore, we adopted two simplification procedures.

First, we truncated all the values of the fluorescence towards zero, and we calculated the average of the results for each integer weight (Table 2).

After that, we grouped the weights in sets with a 4 kDa interval, and the value that represents the set is the maximum value in the range (Table 3). We

did this due to the granularity of the measured molecular weights. In this case, the fluorescence values are very close in the neighborhood, and the deployed hardware may not be as accurate, possibly generating lags of some few kDa. The maximum value in the range was used to represent the interval. It helped not to create false peaks, what could be the case if we have used the sum of the values in the ranges, and not to flat some peaks, what could happen if we have used the average value of the ranges.

Table 2: Examples of data set entries after the fluorescence's values normalization.

Sample ID	10.0	10.1	10.2	...	120.3	120.8	121.0	Health state
d1122	0.0	2.461	4.043	...	24.857	24.709	24.581	Healthy
d1127	33.766	29.228	24.848	...	4.513	4.018	3.284	Healthy
d1132	22.624	21.746	20.784	...	21.093	20.834	20.392	Healthy
d52	6.176	5.156	4.240	...	16.556	16.485	16.485	Diabetes
d56	6.236	5.945	5.775	...	13.592	13.782	13.782	Diabetes
d59	63.151	49.785	36.443	...	11.693	11.629	11.629	Diabetes

Table 3: Examples of data set entries after grouping molecular weights.

Sample ID	10-13	14-17	18-21	...	110-113	114-117	118-121	Health state
d1122	9.455	49.811	24.317	...	24.882	25.244	25.003	Healthy
d1127	12.523	81.814	5.542	...	2.671	3.394	3.915	Healthy
d1132	17.075	28.868	6.731	...	19.325	21.037	20.976	Healthy
d52	5.461	173.141	18.129	...	17.530	16.543	16.609	Diabetes
d56	11.512	44.826	11.464	...	11.888	12.236	13.235	Diabetes
d59	26.700	119.878	27.789	...	9.265	8.962	11.186	Diabetes

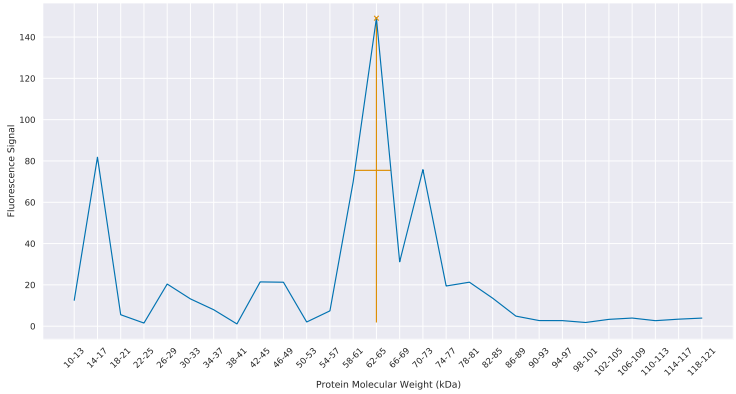
2.4 Peak detection

After the data set was pre-processed, we carried out a peak detection strategy. A peak or local maximum is defined as any entry whose two direct neighbors have a smaller value. Various parameters like prominence, width, and height can be used as thresholds to select specific types of peaks.

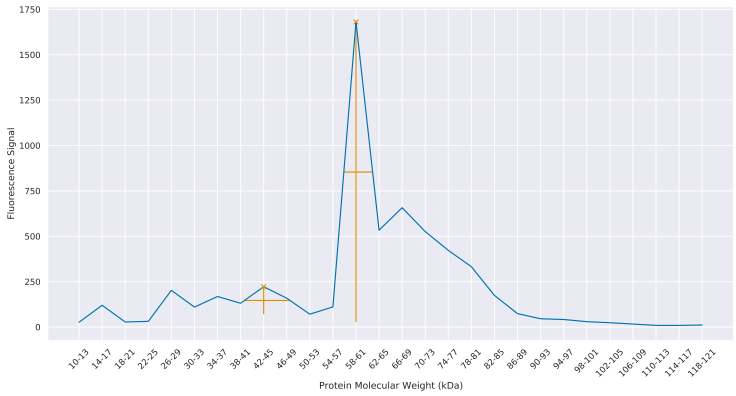
We used an algorithm to automatically detect peaks over each sample using a height threshold of 100. We chose this value because it approximates the average amplitude of the relevant lower peaks in unhealthy individuals, as seen in Figure 1.

With all the peaks detected for all individuals, a new data set is generated. The resulting table contains every height of the distinct peaks detected through

the process as features for every sample, 9 in total. If an individual does not present a specific peak, the height for that will be considered 0 (Table 4).



(a) Healthy individual



(b) Unhealthy individual.

Fig. 2: Graphical representation of peaks detection over samples from Table 3. The "x" axis represents the grouped molecular weights labels. The "y" axis represents the maximum fluorescence value in the molecular weight group. The vertical lines represent the peak prominence and the horizontal ones represent the average width.

2.5 Pairwise relationships

We performed pairwise relationships to identify relationships between the molecular weights of the identified peaks and also the influence that each of them has on the classification of individuals.

Table 4: Examples of peaks data set entries.

Sample ID	26-29	34-37	42-45	46-49	58-61	62-65	66-69	70-73	74-77	Health state
d1122	0.000	0.000	0.000	0.000	0.000	275.493	0.000	0.000	0.000	Healthy
d1127	0.000	0.000	0.000	0.000	0.000	0.000	0.000	149.120	0.000	Healthy
d1132	161.880	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Healthy
d52	358.429	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Diabetes
d56	422.497	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Diabetes
d59	1680.983	0.000	0.000	0.000	0.000	0.000	0.000	222.072	0.000	Diabetes

Pairwise relationships can be understood as any process of comparing entities in pairs to judge which of each entity is preferred or has a more significant amount of some quantitative property, or whether or not the two entities are identical [5].

2.6 Hierarchical clustering

We performed the hierarchical clustering to identify if the given characteristics extracted were sufficient to generate a grouping by the individuals' health status.

Hierarchical clustering is a type of unsupervised machine learning algorithm used to cluster unlabeled data points, grouping the data points with similar characteristics [9]. The calculation that defines the similarity between data points can be different depending on the type of data and how you want to do the grouping.

Because of this grouping property, and because we were looking to explore the data structure to understand emerging profiles, we used hierarchical clustering with the following similarity calculation methods:

1. **average**: Uses the average of the distances of each observation of the two sets.
2. **complete**: Uses the maximum distances between all observations of the two sets.
3. **single**: Uses the minimum of the distances between all observations of the two sets.
4. **ward**: Minimizes the sum of squared differences within all clusters. It is a variance-minimizing approach [20].

The peak heights were treated as coordinates of Euclidean space to calculate the distances.

Another study made over the hierarchical clustering was the silhouette analysis [16]. It consists of calculating the Silhouette Coefficient. It uses the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each given sample. The coefficient for a sample is $(b - a)/max(a, b)$. To clarify, "b" is the distance between a sample and the nearest cluster that the sample is not part of. Silhouette coefficients near 1 indicate that the sample is far away from the neighboring clusters. Values around 0 indicate that the sample is on or very

close to the decision boundary between two neighboring clusters. Negative values indicate that those samples might have been assigned to the wrong cluster.

3 Results

Figure 3 is a graphical representation of the resulting pairwise relationship, a grid of axes such that each variable will be shared in the "y" axis across a single row and the "x" axis across a single column. The diagonal is treated differently, drawing a plot to show the univariate distribution of the data for the variable in that column.

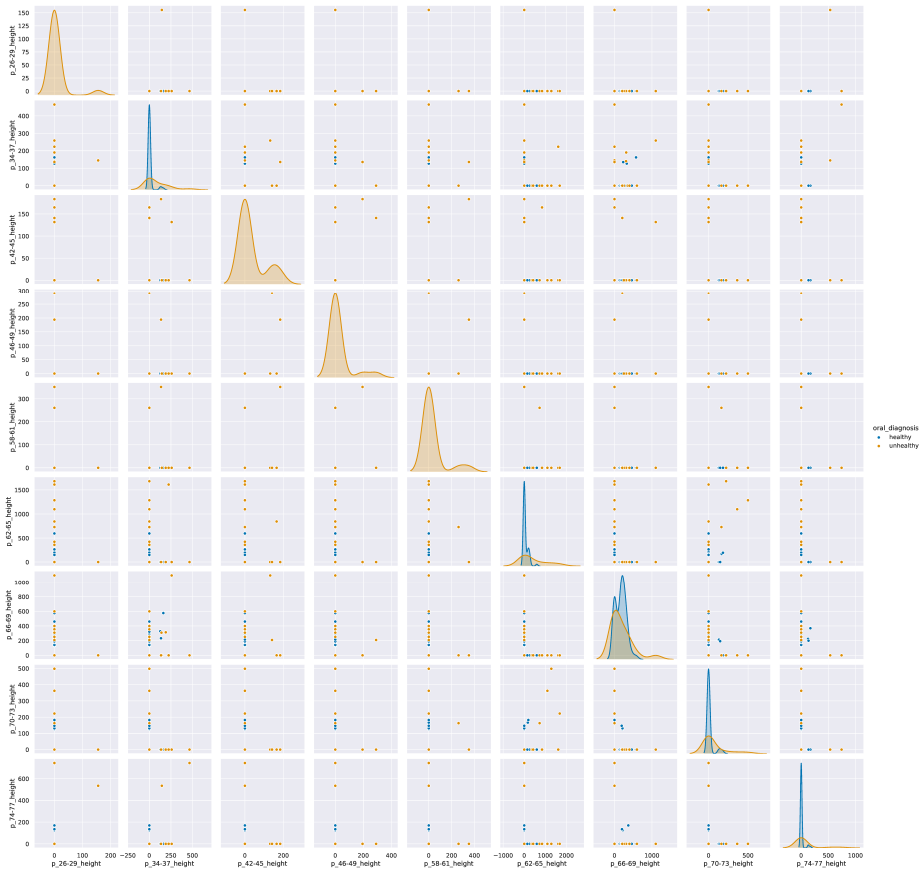
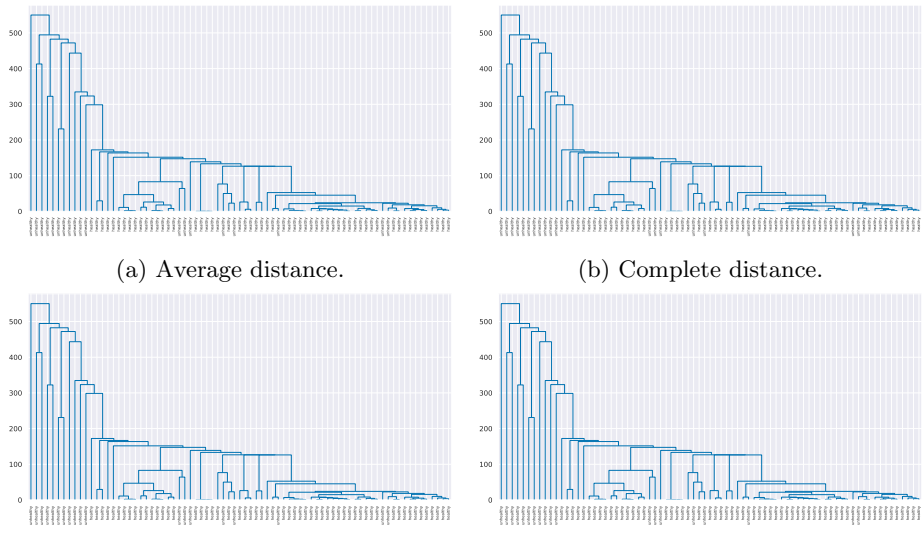


Fig. 3: Pairwise relationships plot of the heights of the peaks to the patient health. Blue represents healthy individuals, while orange represents unhealthy individuals. The graphs in the diagonal axis represent the distribution of individuals for each peak, the remaining represent a binary combination of two peaks in "x" and "y" axis trying to separate the samples.

This representation shows that the peaks “26-29”, “42-45”, “46-49,” and “58-61” are present practically only in unhealthy individuals, making them good candidates for use in the process of differentiation. Furthermore, it shows that in the peaks “34-37”, “62-65”, “70-73,” and “74-77,” healthy individuals are concentrated in lower height values, while the unhealthy individuals are better distributed. Finally, it is noticed that the peak “42-45” has a very similar distribution for both profiles, making it not an exciting feature to be used to classify the health of the base entries.

Also, the other graphs show that no binary combination of attributes could separate healthy and unhealthy individuals well. It indicates that the nature of the attributes requires three or more features to classify the health state of the presented examples.

The hierarchical clustering results, although different distance calculation types were applied, were very similar, presenting a classification of almost all the unhealthy individuals right at the beginning of the formation of the groups, around distance 250, meaning that this is a reasonable distance for the separation of the categories. Figure 4 depicts a graphical representation of the hierarchical clustering performed using every cited method of calculation.



(a) Average distance.

(b) Complete distance.

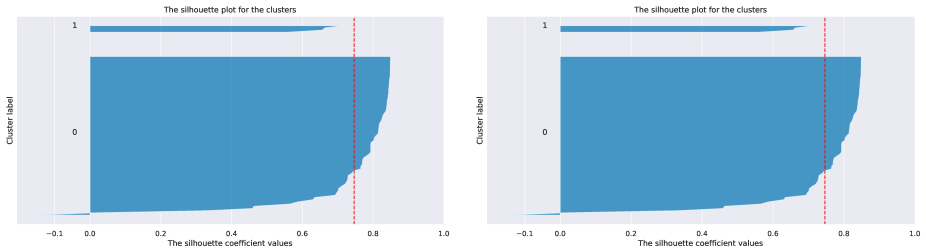
(c) Single distance.

(d) Ward distance.

Fig. 4: Graphical representation of the hierarchical clustering using different methods of distance calculation. On the “x” axis, we have the classifications’ final values, the leaves of the tree, while on the “y” axis, we have the values of the calculated distances.

The rapid agglomeration of unhealthy individuals early in the groupings indicates what the pairwise plot already showed. The features used (peaks heights) manage to characterize well diabetic individuals. Also, as diabetes is a disease with many associated complications, the minority presence of diabetic individuals scattered in other groups may be evidencing the heterogeneity of phenotypes that characterize diabetic patients.

Figure 5 shows the graphical representation of the silhouette analysis. The areas next to the labels "1" and "0" in the "y" axis represent the samples clumping together in cluster "unhealthy" and "healthy," respectively. The dashed lines mark the silhouette coefficient, which is 0.746 for average, complete, and ward methods, showing that they have a reasonable separation distance for the individuals. The coefficient value for the single method is only 0.592.



(a) Average distance. Silhouette Score: 0.746

(b) Complete distance. Silhouette Score: 0.746

(c) Single distance. Silhouette Score: 0.592

(d) Ward distance. Silhouette Score: 0.746

Fig. 5: Graphical representation of the silhouette analysis in the first level of the hierarchical clusterization, with only two clusters. The label "1" represents the unhealthy group and "0" the healthy group. The dashed vertical line is the silhouette score marker.

4 Future Work

As it could be seen in the analyzes made, the data obtained is promising as input for a learning algorithm. Some computational intelligence algorithms are

being tested on this basis to identify an individual health state, the results of that study will be presented in a future work. With positive results in this next study, it would be possible to create an automation for the diagnosis process from the data extracted from a person's saliva using the Experion™Automated Electrophoresis System.

Also, a base with a more significant number of individuals with a greater diversity of diseases is currently being set up. Once this base is ready, we will have more statistical confidence as to whether the peak patterns identified with the current database are sufficient for the health classification and discover new ones if they exist. Besides, we will be able to increase the range of possible classifications, differentiating individuals between healthy and unhealthy (with diabetes), and identifying the specific illness.

5 Conclusions

This article presents an analysis of the saliva protein profiles of diabetic and healthy individuals. The study identified characteristic patterns of variations in the number of specific proteins for these individuals' classification. It indicates that it is possible to quickly and consistently implement a computational intelligence algorithm that can identify a person's health status and automate or assist in a diagnostic process.

The database is limited concerning the number of individuals and the variety of diseases presented. However, the results presented indicate differentiating characteristics between the groups. It is possible to extract these characteristics in a simple way to use them in the process of classification.

6 Ackownledgments

Thanks are due to FCT/MCTES, for the financial support of the Center for Interdisciplinary Research in Health (UID/MULTI/4279/2019). Thanks are also due to FCT and UCP for the CEEC institutional financing of AC Esteves. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

1. Arrais, J.P., Rosa, N., Melo, J., Coelho, E.D., Amaral, D., Correia, M.J., Barros, M., Oliveira, J.L.: Oralcard: a bioinformatic tool for the study of oral proteome. Archives of oral biology **58**(7), 762–772 (2013)
2. Castagnola, M., Scarano, E., Passali, G., Messana, I., Cabras, T., Iavarone, F., Di Cintio, G., Fiorita, A., De Corso, E., Paludetti, G.: Salivary biomarkers and proteomics: future diagnostic and clinical utilities. Acta Otorhinolaryngologica Italica **37**(2), 94 (2017)
3. Conde, J., de la Fuente, J.M., Baptista, P.V.: Rna quantification using gold nanoprobes-application to cancer diagnostics. Journal of nanobiotechnology **8**(1), 5 (2010)

4. Cruz, I., Esteves, E., Fernandes, M., Rosa, N., Correia, M.J., Arrais, J.P., Barros, M.: Salivaprint toolkit–protein profile evaluation and phenotype stratification. Journal of proteomics **171**, 81–86 (2018)
5. David, H.A.: The method of paired comparisons, vol. 12. London (1963)
6. Domingos, P.: A few useful things to know about machine learning. Communications of the ACM **55**(10), 78–87 (2012)
7. Esteves., E., Cruz., I., Esteves., A.C., Barros., M., Rosa., N.: Salivaprint as a non-invasive diagnostic tool. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5: HEALTHINF,. pp. 677–682. INSTICC, SciTePress (2020). https://doi.org/10.5220/0009163506770682
8. Ferreira, A.V., Bastos Filho, C.J., Lins, A.J.: An unsupervised analysis of an alzheimer's disease patient population using subspace search and hierarchical density-based clustering. In: 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI). pp. 1–6. IEEE (2019)
9. Frigui, H., Krishnapuram, R.: Clustering by competitive agglomeration. Pattern recognition **30**(7), 1109–1119 (1997)
10. Kaczor-Urbanowicz, K.E., Martin Carreras-Presas, C., Aro, K., Tu, M., Garcia-Godoy, F., Wong, D.T.: Saliva diagnostics–current views and directions. Experimental Biology and Medicine **242**(5), 459–472 (2017)
11. Kaushik, A., Mujawar, M.A.: Point of care sensing devices: better care for everyone (2018)
12. Lins, A., Muniz, M., Bastos-Filho, C.J.: Comparing machine learning techniques for dementia diagnosis. In: 2018 IEEE Latin American Conference on Computational Intelligence (LA-CCI). pp. 1–6. IEEE (2018)
13. Lins, A., Muniz, M., Garcia, A., Gomes, A., Cabral, R., Bastos-Filho, C.J.: Using artificial neural networks to select the parameters for the prognostic of mild cognitive impairment and dementia in elderly individuals. Computer methods and programs in biomedicine **152**, 93–104 (2017)
14. Loo, J., Yan, W., Ramachandran, P., Wong, D.: Comparative human salivary and plasma proteomes. Journal of dental research **89**(10), 1016–1023 (2010)
15. Rosa, N., Correia, M.J., Arrais, J.P., Lopes, P., Melo, J., Oliveira, J.L., Barros, M.: From the salivary proteome to the oralome: comprehensive molecular oral biology. Archives of oral biology **57**(7), 853–864 (2012)
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
17. Sabbagh, B., Mindt, S., Neumaier, M., Findeisen, P.: Clinical applications of ms-based protein quantification. PROTEOMICS–Clinical Applications **10**(4), 323–345 (2016)
18. Uddin, S., Khan, A., Hossain, M.E., Moni, M.A.: Comparing different supervised machine learning algorithms for disease prediction. BMC Medical Informatics and Decision Making **19**(1), 1–16 (2019)
19. Wang, X., Kaczor-Urbanowicz, K.E., Wong, D.T.: Salivary biomarkers in cancer detection. Medical Oncology **34**(1), 7 (2017)
20. Ward Jr, J.H.: Hierarchical grouping to optimize an objective function. Journal of the American statistical association **58**(301), 236–244 (1963)