Multi-view learning and deep learning for heterogeneous biological data to maintain oral health
Imangaliyev, Sultan

Citation for published version (APA):
Imangaliyev, S. K. (2016). Multi-view learning and deep learning for heterogeneous biological data to maintain oral health

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.
Chapter 9

Epilogue

All models are wrong, but some are useful.

GEORGE E. P. BOX (1919 - 2013)

9.1 Machine learning for oral health: a biological perspective

In recent decades, digital technologies are gradually adopted by dentists and oral health professionals. Many of the routine manual processes are being automated, e.g. communication with patients and processing patient dossiers are done via specialized software and diagnostics is often performed using digital radiology or photography. Readiness to accept and to use these and many other digital technologies by dentists varies largely and depends on many factors. Oral health professionals should see advantages of digital technology based methods over the traditional ones and they should see their positive influence on a treatment quality to accept the technology in their daily practice [213].

Since machine learning is a new technology, it will take a while before dentists realize all possible advantages of using it in their daily practice. This PhD thesis provides an overview of machine learning techniques applied to oral health problems and it is already clear that machine learning can help oral healthcare professionals to solve some of their daily problems in a more efficient and effective way. For example, the algorithm used in Chapter 2 allows to make efficient predictions of an individual’s microbial composition having only a little amount of labeled data and a large amount of unlabeled data. Other algorithms required collecting a large amount of labeled data, but this is not easy to achieve in real world situations. That is why proposed algorithm is more preferred to be used in practical applications. Data clustering algorithms used in Chapter 3 and Chapter 4 illustrate the attempts to show advantage of a more stratified approach to oral health. Data clustering can help to generate new insights to patients health status from data by discovering subgroups of individuals that have similar profiles [159]. These subgroups can subsequently be associated with ecosystem shifts and help to diagnose potential health problems at the early stage. For example, data clustering may reveal groups which have different risk of developing periodontitis in future, e.g. ”starting to be sick” group versus ”might become sick” group versus ”no risk at all for
getting sick” group. Compared to univariate biomarker selection models, the model used in Chapter 4 is explicitly multivariate and combines several markers of health. This may help to increase both optimal sensitivity and specificity of a diagnostic test based on the selected biomarkers.

Integration of salivary metagenomics, metabolomics and biochemistry data in Chapter 6 provided novel insights into healthy oral ecosystem. We propose that the five salivary microbiota clusters identified in these healthy individuals represent different ecological states or 'ecotypes’ of the oral ecosystem, with different ecological properties and different levels of specialization. We find clusters that are functionally adapted to proteolysis and amino acid fermentation, as well as clusters functionally adapted for saccharolysis. Based on our findings, a proteolytic adaptation is reflected in a higher dominance of proteolytic bacteria, and a higher resting pH. Cluster SC3 appears to display signs of advanced ecological specialization towards protein and amino acid metabolism, indicative of a dysbiotic shift towards an early inflammatory state. Striking observation for cluster SC3 is the expensive microbiota community network linkage, the higher levels of the blood-plasma derived albumin, low levels of lysozyme and high levels of chitinase. On the other side of the spectrum, we find an ecological state that is more adept to saccharolytic functions, represented by clusters SC2 and SC1.2. These states are dominated by streptococcal species. Even in a resting state and in prolonged absence of external carbohydrate sources as after an overnight fasting the salivary pH in these individuals is lower compared to the clusters adapted to proteolysis. Interestingly, the saccharolytic cluster SC1.2 showed a marked reduction in species diversity with the lowest number of ecological network connections, as well as a relatively high lysozyme level. We propose that cluster SC1.2 represents an early phase of acidogenic adaptation, specialized in fast sugar metabolism and prone to a more cariogenic state [196]. These findings may indicate that oral health is a result of complex interactions involving the interplay between diet, lifestyle, natural microbiota and immune system, eventually resulting in shifts between different ecological states.

Many oral diseases are caused by interactions between the host and dental plaque [136]. A correct assessment of dental plaque aids the dental experts in the appropriate prognosis of a patient’s gingival condition and their commitment to an oral hygiene programme [156]. Traditionally, dental plaque can be visualised in the clinic by physically touching it with an instrument or by adding a dye. A novel way to look at this plaque is the use of a Quantitative Light-induced Fluorescence (QLF) images [212]. In case of a very large number of images, an automated plaque scoring system presented in Chapter 7 can help dental professionals to avoid personal bias and greatly reduce evaluation error caused by manual plaque assessment. Such an objective system can help to perform more objective profiling of patient risk groups and reduce costs of appropriate preliminary training of dental professionals to distinguish plaque levels. The proposed model can be implemented as user-friendly software for dental practitioners in daily practice and thus contribute to the improvement of patients’ oral health.

In Chapter 8, the associations between red fluorescent plaque and microbial data were evaluated. Results suggested that it is possible to predict an individual’s plaque accumulation level 14 days before the actual plaque measurement. Traditionally, in dental care the patient population is treated as if patients are all at risk for developing disease, hence patients are asked to brush their teeth twice a day, to limit carbohydrate consumption frequency etc. This is caused due to the lack of good prediction models. Currently used risk prediction
models rely on symptoms of a present or previous dental diseases, e.g. caries, gingivitis, periodontitis etc. The model used in that chapter predicts the response to a disease challenge based on biomarkers in the healthy situation. That is why, conceptually, the computational method presented in that chapter is more practical and beneficial in a way how to approach prognosis of a disease development.

9.2 Machine learning for oral health: a mathematical perspective

Machine learning is a very broad field of science and it has many algorithms to solve problems related to oral health domain. Depending on a task, each of the algorithms fits well to the problem and provides an optimal solution. For instance, one of the algorithms used in this thesis to integrate oral health data is a co-regularization algorithm. In the co-regularization setting, several predictor functions are defined over different views of the data, and trained alternately to maximize mutual agreement [185, 186, 109, 110]. Co-regularization approach was used in semi-supervised learning of the oral microbial data in Chapter 2 and multi-view unsupervised learning of the oral microbial data in Chapter 4.

Unlike clinical studies which compare healthy and unhealthy population, studies including only healthy participants pose an additional challenge, because labels are not known. The main hypothesis of the oral health project is that individuals vary in their biological parameters and data clustering approach should be used to identify subgroups of individuals. In recent years, spectral clustering has become one of the most popular modern clustering algorithms. It is simple to implement and very often outperforms traditional clustering algorithms such as the k-means algorithm. The success of spectral clustering is mainly based on the fact that it does not make any assumptions on the form of the clusters. As opposed to k-means, where the resulting clusters are always convex sets, spectral clustering can solve very general problems like intertwined spirals [222]. Spectral clustering was used in single-view learning of the oral metagenomics data in Chapter 3; multi-view learning of the oral metagenomics data in Chapter 4; and single-view learning of the oral metagenomics, metabolomics, and oral biochemistry data in Chapter 6.

One of the biological advantages of the oral cavity is the fact that saliva is a medium easily accessible without any invasion, which may contain biomarkers of oral health. The current development of diagnostic biomarkers via proteomic and genomic techniques in conjunction with technological developments in salivary diagnostics can lead to the development of robust diagnostic tools used by dentists in making clinical decisions and predicting treatment outcomes [232]. From machine learning point of view, biomarker section task is a feature selection task applied to biological data. In this thesis applications are demonstrated for the shrinkage feature selection techniques based on a model regularization such as LASSO [199] or Elastic Net [245]. Regularized unsupervised multi-view feature selection was applied to the oral metagenomics data in Chapter 4; regularized supervised single-view feature selection was applied to the oral metagenomics, metabolomics, and oral biochemistry data in Chapter 6; and only to the oral metagenomics data in in Chapter 8.

One of the problems in the modern oral healthcare approach is an absence of good prediction models, which can identify currently healthy individuals who are at a higher risk of
disease development. Many dental diseases are caused by interactions between the host and dental plaque, thus assessment of the plaque accumulation level can be used to identify individuals at risk. A novel way for visual plaque level assessment is the use of QLF-images [212]. Applying an efficient supervised machine learning algorithm to these images can help dental practitioners to perform more efficient plaque assessments, and provide reproducible, objective, high-throughput computer-aided guidance in identification of patients with high risk. Nowadays, the most efficient image classification models are based on convolutional neural networks [107], because they explicitly learn spatial information between the pixels of an image. Deep convolutional neural network model was applied to the dental QLF-images in Chapter 7.

9.3 Machine learning for life sciences: a future perspective

The beauty of mathematics is in its abstract nature, because it can be applied to any field of science as long as the studied phenomena are quantifiable. Machine learning as a branch of computer science and mathematics inherits this versatile property. Although this PhD thesis is focused on machine learning applications to oral health domain, machine learning can, in principle, help to solve problems in any biomedical domain because of the progress made in a development of multi-omics technologies and computer science.

Multi-omics data plus data from social networks, wearable devices, sensors, mobile apps and many other sources can be stored and analyzed using sophisticated machine learning algorithms. This can help to make healthcare system better by diverging from a simplistic one size fits all approach to precision medicine, which is different from personalized medicine. Personalized medicine implies that unique treatments can be designed for each individual, which is practically impossible to reach at this time, because of technical and economic limitations. Instead of that, precision medicine means the ability to split individuals into groups that differ biologically with respect to susceptibility to disease, disease development or response to treatment. Precision medicine and big data-enabled medicine have been receiving increased attention in the scientific community due to its high complexity and ambitious challenges [34]. Some experts suggest to apply data integration strategies used in Geographical Information Systems (GIS) to omics-based medicine data, introducing the term The Human GIS [201]. The idea of combining all possible data to improve diagnostics and treatment of diseases led to the launch of the Precision Medicine Initiative by the US president Barack Obama, who announced it during his State of the Union Address in January 20, 2015 [42].

Of course, it is possible that our society is now experiencing a peak of inflated expectations from machine learning and big data. However, even in that case one must not become cynical and give up on his way to seek the truth. The progress of computer science should help to deliver better medical treatment and bring research results closer to the medical practitioners, and thus to real patients suffering from diseases such as cancer, HIV, diabetes mellitus etc. The history of computer science itself is a good example that it is possible, because in the 50s of the previous century computers were big, slow and they were used exclusively for computing. However, due to technological progress made in the second half of the 20th century, computers, smartphones and tablets are nowadays an essential part of our everyday life. Similarly to this, machine learning can transform our approach in life sciences and help
us to build a healthier society. This is something I hope to actively contribute to during the forthcoming years as a scientist.