Machine Learning for Salivary Gland Ultrasound Scoring

Section 304

Preliminary Deliverables

10/11/2023

Team:

Richard Yang (tyang296@wisc.edu)	Team Lead
Yousef Gadalla (ygadalla@wisc.edu)	Communicator
Brandon Drew (bsdrew2@wisc.edu)	BSAC
Dhruv Nadkarni (dnadkarni@wisc.edu)	BWIG
Siya Mahajan (mahajan24@wisc.edu)	BWIG
Aran Viswanath (viswanath3@wisc.edu)	BPAG

Abstract

Sjögren's syndrome (SjS) is a systemic autoimmune disease (SAD) that causes dysfunction of the exocrine glands (mainly the salivary and lacrimal glands) with patients often showing persistent dryness of the mouth and eyes [1, 2]. The current standard of care of the client is to perform at least baseline salivary gland ultrasounds (of the submandibular and parotid glands) in patients who potentially have SjS. For some higher-risk individuals, regularly scheduled salivary gland ultrasounds are performed. The problem arises within the current Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) ultrasound grading system, which requires subjective opinions and lacks nuance. As a result, a machine learning approach is proposed to reduce inter-reader variability and to provide a more exact disgnosis. The proposed algorithm takes ultrasound grayscale images as input and outputs SjS positive or SjS negative. The team has proposed a K Nearest Neighbour (KNN) model to assess what performance can be reasonably expected from models built on the same dataset and VGG-19 for final optimization. The results will be summarized in accuracy results, confusions matrices, and Receiver Opearting Charateristic (ROC) curves. The performance of the final model will be assessed relative to the baseline model.

Table of Contents

Introduction

Sjögren's syndrome (SjS) is a condition estimated to affect a significant population, ranging between 1 and 4 million individuals in the United States [3]. Typically, patients are diagnosed after the age of 50, with a noticeable prevalence among women [4]. While SjS currently lacks a cure, treatment options exist, tailored to the specific affected areas. Obtaining a precise and swift diagnosis with minimal invasiveness is crucial. Such a diagnosis plays a vital role in ensuring timely and suitable medical intervention, thereby reducing the associated risks of trauma, infection, and recovery.

Currently, a number of diagnostic methods are employed to detect SjS, including the Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) Ultrasound Scoring System, blood and urine tests, Schirmer tear tests, Sialography, and Lip Biopsies. While these methods exhibit efficacy in SjS detection, they each present distinct challenges pertaining to accuracy, speed, and invasiveness.

The OMERACT Ultrasound Scoring System encompasses a set of guidelines used for the interpretation of ultrasound images of the parotid and submandibular glands [5]. While this approach minimizes invasiveness, it relies on human interpretation, introducing subjectivity into the diagnostic process and potentially causing delays.

A machine learning model to detect SjS from ultrasound images is desirable as it enhances the noninvasive procedure of taking ultrasound scans by removing human subjectivity and allows for quicker diagnosis resulting in less strain on clinical staff and quicker access to treatment for patients.

Background

Sjögren's syndrome (SjS) is an autoimmune disorder that is generally characterized by two main symptoms: dryness of the eyes and dryness of the mouth. This results in the manifestation of its most common complications: dental cavities, yeast infections, and vision problems and is often found in conjunction with other rheumatic diseases [6].

The OMERACT is a scoring system that utilizes salivary gland ultrasounds to diagnose SjS and other rheumatic diseases. It is characterized by a four-grade scoring system based on the parotid and submandibular glands in patients starting at 0, normal appearance, and going to 3, maximum change from the norm. Despite its popularity, it is still not present in the 2016 American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) classification criteria so it is often used as an initial step to determine if a patient is not at risk for SjS or if other tests should be performed to determine if the patient does or does not have SjS [7].

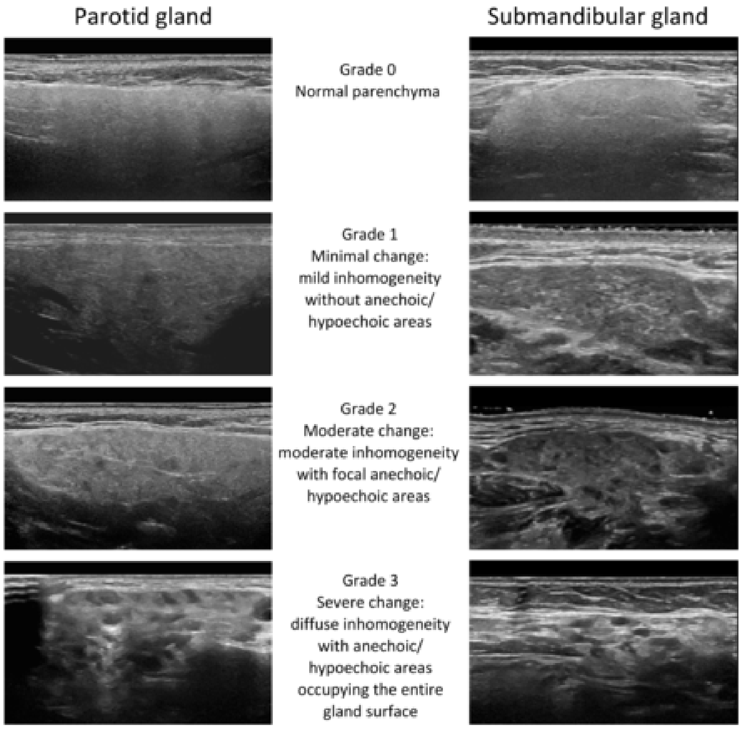

Figure 1: OMERACT grading system [5]

Conversely, the other methods, while capable of diagnosing SjS, introduce invasive complications and accuracy issues. Blood and urine tests, for instance, may be susceptible to sample contamination, which could lead to unnecessary hospitalization and the exposure of patients to unwarranted medications, thereby posing both safety and financial risks [8].

The Schirmer Tear Test, which entails the insertion of a filter paper strip into the patient's eyelid to measure tear travel distance, can cause discomfort and the potential for infection due to the foreign body insertion [9]. Sialography, another imaging method, should be used sparingly, given its requirement for patient sedation, contrast dye injection into the salivary glands, and radiation exposure through X-rays. This procedure carries risks of salivary duct damage, swelling, and tenderness [10].

Lastly, the inner lip biopsy, involving the extraction and analysis of lip tissue, represents a significantly time-consuming and invasive procedure compared to an ultrasound scan. It necessitates surgery for tissue removal, followed by the separation and transfer of glands to a pathologist with specialized training for SjS diagnosis [11]. Consequently, this process may result in treatment delays until confirmation from the pathologist is obtained.

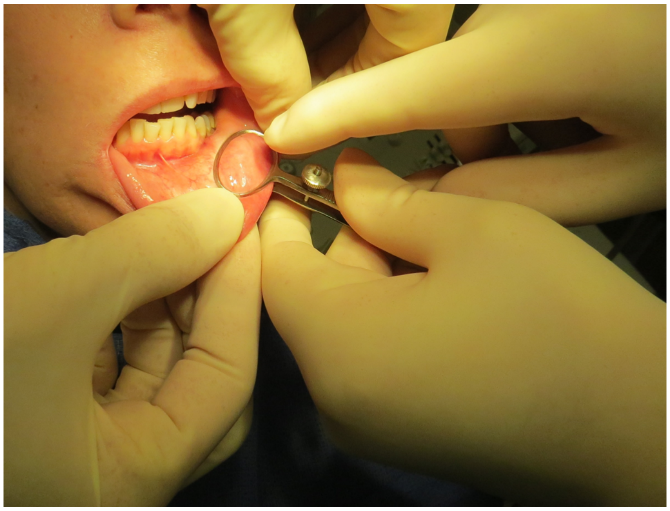

Figure 2: lip biopsy [12]

The client, Dr. Sara McCoy, is looking for a machine learning algorithm that will eliminate subjectivity and be based on the OMERACT system to diagnose SjS when input with ultrasound pictures of a patient's salivary glands.

The algorithm will be input with de-identified pictures of a patient's salivary glands and output a binary result: positive or negative, as it relates to a patient's proposed SjS diagnosis. There will be a baseline model and a final model both of which will be trained and tested with the same dataset split 7-3 training-validation.

Preliminary Designs

Baseline Algorithms

To evaluate the effectiveness of the final algorithm design, the team opted to also write an algorithm for a simplified machine learning program that will act as a baseline. This baseline will be utilized for comparison and ensure that the accuracy of the final algorithm is significantly better than that of the baseline. All of these algorithms will be examples of supervised learning, meaning that they will be provided data that is already classified and analyze these data points to create their generalizations/predictive capabilities for new data sets. Four algorithms were considered for the baseline model: support vector machine (SVM), K-nearest neighbor (KNN), random forest, and import vector machine (IVM).

The support vector machine is an algorithm that performs binary classification. The algorithm takes in the data, and based upon the varying characteristics being observed, creates a hyperplane that separates the categories [13]. This hyperplane is then used to predict the classification of new data points. A limitation of the SVM is its usage with non-linear datasets, which make hyperplane creation more difficult and inaccurate [14].

The K-nearest neighbor algorithm (KNN) is an algorithm that places the training data upon a grid based upon some empirical value. When new data is added, the distance between this new data point and existing data is calculated. Then based upon the known classifications of the K nearest data points, this new data point is assigned the value of the average classification [15]. For example, given K=1, the closest existing data point will be used to predict the diagnosis of the new data point. With a K=3, the average diagnosis of the three closest neighbors would be used as the diagnosis of the new data point.

The random forest algorithm utilizes a series of classifications/data comparisons with weights that will be used in the final calculation of the diagnosis [16]. The tree is typically created by utilizing the training data and random attributes are taken from the data to create a tree diagram with nodes and leaves. Given the importance of a given characteristic, a weight will be assigned to the characteristic to reflect said importance. As new data is added, the data will be placed into the tree and be compared to the existing branches[16]. At the end, the score given by the tree will provide the diagnosis. To improve the tree's accuracy, the process of making trees can be repeated to form several coexisting trees. All of these trees would then output some score, and the average diagnosis will be used to characterize the new data [16].

The import vector machine (IVM) is an updated version of the SVM that has been developed to improve upon the SVM's limitation in nonlinear/complex data classification [17, 18]. This algorithm's central idea is the transformation of data utilizing a kernel method. This allows for the data to be manipulated to create a more linear relationship that allows for a hyperplane to be created, even when the original data are nonlinear.

Final Algorithms

For the final algorithm, five supervised learning algorithms were considered: ResNet-50, deep neural networks (DNN), convolutional neural networks (CNN), VGG-19 and U-net.

ResNet-50 is a machine learning algorithm with 50 layers. The algorithm uses these layers to analyze data to better characterize and separate the data. This algorithm utilizes a method known as "skip connections" that allows it to ignore specific layers that have been found to damage the framework and accuracy of the model [19]. This allows the algorithm to learn from itself and continually improve with more data sets.

DNNs are an algorithm type that is characterized by how it can analyze data and isolate important predictive features for the data [20]. This is done by having many layers that analyze the data and perform small transformations to the inputted data, which is done to mimic the brain's way of processing information.

CNNs are a subset of DNN that is specifically designed for image processing [21]. CNN's have many different layer types to fulfill the algorithm's task of image classification: the first layer is a convolutional layer, which extracts physical information from the images. This data is then put through the pooling layers, which simplifies the images and holds the most important data features [22]. Finally this is analyzed by connected layers that analyze the simplified image and data features to draw predictions from the data.

VGG-19 is a type of CNN and DNN with 19 pre existing layers that are pre- trained [23]. The high amount of layers allows this algorithm to excel at information extraction from data, which helps improve the accuracy of the algorithm [24]. This algorithm is used in many research papers for image classification and analysis [23,24].

UNet is a CNN variant algorithm that is typically applied in processing biomedical images. The necessity of classifying medical images requires UNet to be good at analyzing each pixel of a given image. This helps the algorithm to locate areas of interest within a given image [25]. Once the area of interest has been located, convolutional layers are applied to isolate important features which can be used later for image analysis [25].

Criteria Descriptions

Accuracy is the percentage of correct classifications in relation to total predictions made. This is a very important part of the algorithm, as it has to minimize the percentage of errors. If the algorithm were not accurate, data would be incorrectly labeled, resulting in either a false positive or false negative diagnostic. As a result, patients would either receive unnecessary treatment or not get any treatment at all.

Processing speed is defined as how quickly a computer can process data or instructions. In terms of machine learning, it is how quickly a model processes data, interprets data based on previous data, and produces a predicted output. The processing speed interacts with building complexity and compactness, as complexity is defined by how many layers and filters a program has, which in turn impacts the speed at which a computer can process data. Programs that do not process, interpret, and produce an output, based on the provided Ultrasound images, quickly and efficiently have a lower score.

Building Complexity looks at how complex the algorithm is to code. This is impacted by how many layers that the learning algorithm has for analyzing the data. Programs with a larger amount of layers and filters for analyzing data have lower scores reflecting their complexity.

Compactness looks at the size of the algorithm. A smaller algorithm size is better as it takes less space on a hard drive.

Scalability is the improvement potential of the algorithm with increasing dataset i.e., performance = O(n) where n is the size of the dataset. This encompasses how easily the structure of the algorithm can be adjusted, model be re-trained, the weights be reset, and the flexibility of the hyperparameters. A larger improvement potential is desirable so that the accuracy of the algorithm can scale with more complete training data.

Baseline Design Matrix

	Baseline Model				
	SVM	KNN	Random Forest	IVM	Weights
Accuracy/Safety	14	18	16	16	20
Processing Speed	10.5	10.5	9	13.5	15
Building Complexity	8	9	5	7	10
Compactness	7	8	6	8	10
Scalability	7.5	12	9	12	15
Sum	67.14%	82.14%	64.29%	80.71%	70

Table 1: Baseline Model Design Matrix

Baseline Design Scorings

Scores for each algorithm were determined based upon published data from papers and studies conducted by institutional and research bodies.

The support vector machine scored relatively low in the accuracy category due to SVM's limitation in non-linear data analysis. As the classification of salivary ultrasounds will likely not fall into a neat linear pattern, the SVM received a score of 14/20 for accuracy/safety. The processing speed of SVM's follows a time complexity of O(n^2) [13]. This means as the number of data points increases, the amount of time that the SVM runs will increase quadratically. This is seen in SVM's low processing speed score. The low scalability score of SVM is connected to how the addition of more non-linear data will cause the hyperplane to be more inaccurate, especially as the hyperplane does not typically change with the addition of new data points.

The KNN had similar processing speeds to SVM due to its requirement to draw new connections between each datapoint to existing data. The accuracy is deemed to be higher due to research studies that compared KNN vs SVM algorithms and found improved KNN accuracy within EEG data reading along with other medical data analysis [26]. This is seen in KNN's score of 18/20 for accuracy. The scalability score of KNN is scored relatively high because of KNN's ability to adapt new data into its predictive algorithm for diagnostic purposes.

In a study comparing KNN and random forest, the accuracy of random forest was found to be less than KNN (96% vs 84% accuracy) [27]. This is reflected in the random forest's score of 16/20. Due to the need for random trait selection from the data set and importance scoring of these traits, the building complexity of the random forest is relatively low compared to the other algorithms.

The IVMhas a higher score compared to the SVM due to the IVM being made as an improvement of the SVM in many of the categories. The IVM has been altered to utilize a kernel method to make its accuracy better for the complex data that will likely be present in the ultrasound analysis of salivary glands. Also, as the processing speed is meant to have been improved as well, the processing speed is higher and thus the score is higher as well. The only category that the IVM is worse in is the complexity category because the changes to the SVM to improve it require more complex coding and data transformation.

Final Design Matrix

	Final Model					
	ResNet-50	DNN	CNN	VGG-19	U-net	Weights
Accuracy/Safety	18	18	16	18	16	20
Processing Speed	7.5	9	10.5	10.5	10.5	15
Building Complexity	7	4	6	7	6	10
Compactness	6	6	7	7	7	10
Scalability	10.5	12	10.5	9	10.5	15
Sum	70.00%	70.00%	71.43%	73.57%	71.43%	70

Table 2: Final Model Design Matrix

Final Design Scorings

As many of the algorithms are variations of each other, there are many criteria where they have similar values. The ResNet-50, DNN, and VGG-19 were found to be relatively similar in accuracy due to their ability of multiple layer analysis of images. CNN and U-net were scored slightly lower due to reports of their accuracy decreasing with pixely images [21, 28].

The building complexity of all the algorithms were relatively low, however ResNet-50 and VGG-19 were scored slightly higher due to the pre-training that they have received, which could potentially alleviate the building complexity. This analysis can change down the line as this pre-training could potentially hinder the ultrasound analysis of the algorithms.

Fabrication/Development Process

Materials:

This product will only consist of software. For this reason, there will be no physical materials used when building either the baseline or final algorithm. Therefore, the client will need a computer that has Python installed. As for software, PyTorch will be used for the machine-learning framework, and GitHub will be necessary for maintenance. So overall this product will cost nothing to produce.

Methods

Two models will be used to predict whether or not a patient is Sjögren positive or negative from the ultrasounds. The first model, KNN, will be used strictly for baseline testing. This means the model will assess the baseline performance but will not be optimized for the dataset. The second model, VGG-19, will be used for final validation. This model will be optimized for the dataset but will also have to avoid overfitting the data.

Since the product will be highly adaptive, the overall accuracy of the models will improve with more data. Therefore, it has been created to increase in reliability with time and usage. Both models will be trained and tested by partitioning the dataset into training and validation sets with a 7:3 ratio. The outputs will then be put into a confusion matrix and a ROC curve will be generated. Overall, the accuracy of the final model should be greater than or equal to 95%.

It should be noted that in the early stages of the product, a physician's opinion will likely be needed to supplement the output of the algorithm.

Testing

The team plans on splitting the data with respect to a 7:3 ratio, implying a 70% training and 30% testing split. The main criteria being tested with the respective testing data is accuracy, as the team monitors the accuracy of the models. The test will be conducted with unlabeled data, as the team will un-label the set of testing data. The team will import a file of unlabeled vector data into the model, and the model's classification of the data will be compared to the labeled counterparts. The accuracy will be calculated by the percentage of correct decisions. Using the information, the team will view the incorrectly labeled data and make updates to the models respectively. The other criteria will be tested via comparing the two proposed models, the baseline KNN and final VGG-19. Since the VGG-19 model is the teams final implementation, all criteria will be compared between the two, including accuracy. Using the results from the testing phase, the teams will either update the models or begin working on the frontend implementation.

Discussion

When working with Ultrasound data, it is important to note its challenges: operator dependency, variability of ultrasound machines across different manufacturers, and differential image quality with structures behind bone and air [29]. The ultrasound operator does not only dictate the perceived quality of the images, but also orientation and field-of-view. The same effects can also be induced by the differences between machines. These sources of variability lead to dramatic differences in the acquired images. The testing result generated is therefore only indicative of the performance of the model within the scope of the ultrasound images taken with the same techniques and machines. Thus, intra-disease generalizability cannot be guaranteed. Similarly, inter-disease generalizability can also be affected if another disease requires the imaging of a structure behind bone and air.

Conclusions

In an effort to classify SjS ultrasound images, the team has devised building and testing guidelines for the machine learning algorithm. Furthermore, the team has proposed two models, KNN and VGG-19 for baseline testing and final validation respectively. While the baseline model will be used solely for assessing baseline performance and will not be optimized, the final model will be optimized for the dataset while avoiding overfitting and its performance will be compared against the baseline results. The team has also outlined the procedures for training these models and the matrices for comparing the results.

Future work

In the near future, the team will split into three groups and work on their respective parts, either the baseline KNN model, the final VGG-19 model, or an image processing script. In addition to the given data, the team will also continue searching for more ultrasound salivary gland images for more training and testing to continuously update the model. After all models are complete, the team will work on an easy to use frontend interface for the client. In the far future, the team would update the model with new data and potentially upscale the model to aid with the scoring of other ultrasound images.

References

1. M. Ramos-Casals and J. Font, “Primary sjögren’s syndrome: Current and emergent AETIOPATHOGENIC concepts,” Rheumatology, vol. 44, no. 11, pp. 1354–1367, 2005. doi:10.1093/rheumatology/keh714

2. P. Brito-Zerón, S. Retamozo, and M. Ramos-Casals, “Síndrome de Sjögren,” Medicina Clínica, vol. 160, no. 4, pp. 163–171, 2023. doi:10.1016/j.medcli.2022.10.007

3. “Sjögren’s Disease | National Institute of Dental and Craniofacial Research,” www.nidcr.nih.gov. https://www.google.com/url?q=https://www.nidcr.nih.gov/health-info/sjogrens-disease&sa=D&source=docs&ust=1697072274306517&usg=AOvVaw3KghYeXwXBUyr2h7In5wgy (accessed Oct. 11, 2023).

4. N. S. C. and O. Branch, “Sjögren’s Syndrome,” National Institute of Arthritis and Musculoskeletal and Skin Diseases, Apr. 07, 2017. https://www.niams.nih.gov/health-topics/sjogrens-syndrome/basics/symptoms-causes

5. V. Fana, U. M. Dohn, S. Krabbe, and L. Terslev, “Application of the omeract grey-scale ultrasound scoring system for salivary glands in a single-centre cohort of patients with suspected Sjögren’s syndrome,” RMD Open, vol. 7, no. 2, 2021. doi:10.1136/rmdopen-2020-001516

6. “Sjogren’s syndrome,” Mayo Clinic, https://www.mayoclinic.org/diseases-conditions/sjogrens-syndrome/symptoms-causes/syc-20353216 (accessed Oct. 11, 2023).

7. O. Al Tabaa et al., “Normal salivary gland ultrasonography could rule out the diagnosis of Sjögren’s syndrome in anti-ssa-negative patients with SICCA syndrome,” RMD Open, vol. 7, no. 1, 2021. doi:10.1136/rmdopen-2020-001503

8. “Blood Culture Contamination: An Overview for Infection Control and Antibiotic Stewardship Programs Working with the Clinical Laboratory.” Available: https://www.cdc.gov/antibiotic-use/core-elements/pdfs/fs-bloodculture-508.pdf

9. E. Staff, “Schirmer’s test: A test for dry eyes,” The Eye News, https://theyenews.com/schirmers-test/ (accessed Oct. 11, 2023).

10. A. Delgado, “Salivary duct stones: Causes, symptoms, and diagnosis,” Healthline, https://www.healthline.com/health/salivary-duct-stones#outlook (accessed Oct. 11, 2023).

11. “Labial gland (lip) biopsy,” Johns Hopkins Sjögren’s Center, https://www.hopkinssjogrens.org/disease-information/diagnosis-sjogrens-syndrome/labial-gland-lip-biopsy/ (accessed Oct. 11, 2023).

12. “Iowa head and Neck Protocols,” Lip Biopsy for Sjogren’s Syndrome (Minor Salivary Gland Biopsy) Using Chalazion Clamp | Iowa Head and Neck Protocols, https://medicine.uiowa.edu/iowaprotocols/lip-biopsy-sjogrens-syndrome-minor-salivary-gland-biopsy-using-chalazion-clamp (accessed Oct. 5, 2023).

13. Noble, W. What is a support vector machine?. Nat Biotechnol 24, 1565–1567 (2006). https://doi.org/10.1038/nbt1206-1565

14. D. Westreich, J. Lessler, and M. J. Funk, “Propensity score estimation: Neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression,” Journal of Clinical Epidemiology, vol. 63, no. 8, pp. 826–833, 2010. doi:10.1016/j.jclinepi.2009.11.020

15. Sarkar, M, and T Y Leong. “Application of K-nearest neighbors algorithm on breast cancer diagnosis problem.” Proceedings. AMIA Symposium (2000): 759-63

16. L. Frederick, “Implementation of Breiman’s Random Forest Machine Learning Algorithm,” ECE591Q Machine Learning Journal Paper. Fall 2005.

17. Ji Zhu & Trevor Hastie (2005) Kernel Logistic Regression and the Import Vector Machine, Journal of Computational and Graphical Statistics, 14:1, 185-205, DOI: 10.1198/106186005X25619

18. Suthaharan, S. (2016). Support Vector Machine. In: Machine Learning Models and Algorithms for Big Data Classification. Integrated Series in Information Systems, vol 36. Springer, Boston, MA. https://doi.org/10.1007/978-1-4899-7641-3_9

19. Mukherjee, S. "The Annotated ResNet-50" Toward Data Science. August, 2022. https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

20. A. R. N. Aouichaoui, R. Al, J. Abildskov, and G. Sin, “Comparison of group-contribution and machine learning-based property prediction models with uncertainty quantification,” 31st European Symposium on Computer Aided Process Engineering, pp. 755–760, 2021. doi:10.1016/b978-0-323-88506-5.50118-2

21. Al'Aref, Subhi J et al. “Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging.” European heart journal vol. 40,24 (2019): 1975-1986. doi:10.1093/eurheartj/ehy404

22. J. Raitoharju, “Convolutional Neural Networks,” Deep Learning for Robot Perception and Cognition, pp. 35–69, 2022. doi:10.1016/b978-0-32-385787-1.00008-7

23. Karen Simonyan, Andrew Zisserman: Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015

24. Cao, Zili et al. “BND-VGG-19: A deep learning algorithm for COVID-19 identification utilizing X-ray images.” Knowledge-based systems vol. 258 (2022): 110040. doi:10.1016/j.knosys.2022.110040

25. Zhang, J. "UNet- Line by Line Explanation" Towards Data Science. October 2019. https://towardsdatascience.com/unet-line-by-line-explanation-9b191c76baf5

26. M.N.A.H. Sha’abani, N. Fuad, N. Jamal, M. Ismail, “kNN and SVM Classification for EEG: A Review” InECCE2019 Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, vol. 632, pp. 544-554, 2019.

27. M. Shapiee, M. Ibrahim, M.Razman, M. Abdullah, R. Musa, M. Hassan, A. Majeed, “The Classification of Skateboarding Trick Manoeuvres Through the Integration of Image Processing Techniques and Machine Learning,” InECCE2019 Proceedings of the 5th International Conference on Electrical, Control & Computer Engineering, vol. 632, pp. 544-554, 2019.

28. Yin, X. X., Sun, L., Fu, Y., Lu, R., & Zhang, Y. (2022). U-Net-Based Medical Image Segmentation. Journal of healthcare engineering, 2022, 4189781. https://doi.org/10.1155/2022/4189781

29. L. J. Brattain, B. A. Telfer, M. Dhyani, J. R. Grajo, and A. E. Samir, “Machine learning for medical ultrasound: Status, methods, and future opportunities,” Abdominal Radiology, vol. 43, no. 4, pp. 786–799, 2018. doi:10.1007/s00261-018-1517-0

Appendix

Product Design Specifications

Machine Learning for Salivary Gland Ultrasound Scoring

Section 304

Product Design Specifications

09/22/2023

Team:

Richard Yang (tyang296@wisc.edu)	Team Lead
Yousef Gadalla (ygadalla@wisc.edu)	Communicator
Brandon Drew (bsdrew2@wisc.edu)	BSAC
Dhruv Nadkarni (dnadkarni@wisc.edu)	BWIG
Siya Mahajan (mahajan24@wisc.edu)	BWIG
Aran Viswanath (viswanath3@wisc.edu)	BPAG

UW BME Design

Background

Sjögren's syndrome (SjS) is a systemic autoimmune disease (SAD) that causes dysfunction of the exocrine glands (mainly the salivary and lacrimal glands) with patients often showing persistent dryness of the mouth and eyes [1, 2]. According to estimations, two to four million people in the United States have SjS; however, only one million have been diagnosed, likely due to the nonspecific diagnostic guidelines and the heterogeneous nature of the disease [3]. The current standard of care of the client is to perform at least baseline salivary gland ultrasounds (of the submandibular and parotid glands) in patients who potentially have SjS. For some higher-risk individuals, regularly scheduled salivary gland ultrasounds are performed.

Function

The problem arises within the current Outcome Measures in Rheumatoid Arthritis Clinical Trials (OMERACT) ultrasound grading system, which requires subjective opinions and lacks nuance. As a result, a machine learning approach is proposed to reduce inter-reader variability and to provide a more exact prognosis. The proposed algorithm takes ultrasound grayscale images as input and outputs SjS positive or SjS negative.

Client requirements

The following is a list of client requirements:

- The algorithm needs to take ultrasound grayscale images as input and output binary labels of SjS positive or SjS negative.
- It is preferable that the algorithm can be processed in real-time, such that the physicians can receive the algorithm's output immediately after the patient's ultrasound procedure.
- Images must be de-identified before they can be used for training
- Generalizability to other Rheumatic diseases and Emergency Medical Technician (EMT) applications is preferable

Design requirements

1. Physical and Operational Characteristics

a. Performance requirements:

The product will be a machine learning program that is run on hospital computers and analyzes salivary gland ultrasound images. The program must provide an accurate classification of the images and determine whether the patient has SjS or not.

The program will be utilized in clinical settings post-ultrasound readings. This means that the device could potentially be used many times a day, depending on clinic hours and number of patients that need ultrasounds. To ensure that no long waits occur for patients, the machine learning algorithm should be able to generate results within 15 minutes. A first-in-first-out (FIFO) queue structure will be used to ensure that no tasks are skipped due to processing time.

b. Safety:

As this is a machine learning program, there should not be any safety concerns for users; however, as this algorithm will be utilized in diagnosing SjS, it is very important that the algorithm works properly. Otherwise, any missed diagnosis could result in patient's not receiving proper treatment for SjS, which potentially can cause increased health risks and concerns [4].

c. Accuracy and Reliability:

Since this is a highly adaptable product, it will gain accuracy as it is presented with more data. Thus it will be created to increase in reliability with additional time and usage. The models will be evaluated by first partitioning the dataset into training and validation sets with a 7:3 ratio respectively. The model will then be trained on the training set and evaluated with the validation set. The output of which will be put into confusion matrices and the accuracy results as well as Receiver Operating Chracteristic (ROC) curve will be generated.

A baseline performance (performance of a simple model with the same training data as the final model) will first be assessed using a support vector machine (SVM), and the goal is to perform better than the baseline with either a more complicated deep neural network (DNN) or an established model like the ResNet-50. Ideally, the accuracy should be greater than or equal to 95%.

In practice, especially in the early stages of the product, a physician's opinion might be needed to supplement the output of the algorithm.

d. Life in Service:

In light of a better scoring system, or a software/hardware change this product is not compatible with, this product may become obsolete. As a machine learning algorithm, however; it is can be updated by the team in the future when new data becomes available to improve performance.

e. Shelf Life:

Given that the system is updated in order to stay relevant with the software and hardware it will be run on, the shelf life of this product is infinite.

f. Operating Environment:

The product is designed to operate in clinical environments, primarily on computers that can run the code. The code can run on any operating system but requires Python to be installed on the computer for the program to run if the client prefers the program in a .py or .ipynb format. If the code is built as an executable software, no Python is required.

g. Ergonomics:

The sole restrictions would be the usage of an admissible computer, the requirement of Python dependent upon the client's preferred file format, and patient permission for their images to be run through the program.

h. Size:

As the product is software oriented, there are no physical size restrictions or requirements.

i. Weight:

The project design is software based, and thus weight is not applicable in terms of software. The weight required by the client ranges, as they require a workstation, whether a laptop or desktop, to run the software and process images.

j. Materials:

There only will be a software aspect to the product. So, since there will be no hardware, no physical materials are needed for this product. As for software, PyTorch will be used for the machine learning framework, and GitHub will be necessary for maintenance. Depending on the processing speed of the final model, a GPU module might be required to decrease processing time.

k. Aesthetics, Appearance, and Finish:

There is no hardware, so there will be no color, shape, or form texture requirements. This product consists of only software, so aesthetics, appearance, and finish are not applicable.

2. Production Characteristics

a. Quantity:

Only one program has to be written to fulfill the requirements. This program will then be used on any device the client wishes to use.

b. Target Product Cost:

Since this device only consists of software, there will be no manufacturing costs.

3. Miscellaneous

a. Standards and Specifications:

The project concerns human data; thus, a few issues must be addressed, namely the acquisition of human data, de-identification protocols, and working with de-identified data.

De-identified ultrasound images will be provided by the client; however, if any additional data acquisition is to take place, per 21 CFR 56.102, any data acquisition from human subjects shall fall under the definition of clinical investigation and:

> must meet the requirements for prior submission to the Food and Drug Administration under section 505(i) or 520(g) of the act, or need not meet the requirements for prior submission to the Food and Drug Administration under these sections of the act, but the results of which are intended to be later submitted to, or held for inspection by, the Food and Drug Administration as part of an application for a research or marketing permit. [5]

Human subject shall be defined as an individual who is or becomes a participant in this project, as the subject of ultrasound imaging [5]. In such a case, informed consent of the participants and IRB approval must be obtained. Per FDA guidelines, adequate information that allows an informed decision must be provided, participants' understanding of the aforementioned information should be facilitated, adequate time must be allocated for the participants to ask questiosn and discuss protocols with family and friends, and voluntary participation agreement must be obtained, and the participants should be updated with more information as research progresses [6].

In the case of working with de-identified data, which is defined as there is no reasonable basis to believe that the information can be used to identify an individual under 45 CFR 164.514, HIPAA Privacy Rule "does not restrict the use or disclosure of de-identified health information, as it is no longer considered protected health information" [7, 8].

Per 45 CFR 164.514(b), HIPAA provides two de-identification methods: 1) Expert determination and 2) Safe harbor. The former requires “a person with appropriate knowledge of and experience with generally accepted statistical and scientific principles and methods for rendering information not individually identifiable” while the latter requires the removal of 18 types of identifiers, including but not limited to name, address, and phone number [7].

b. Customer:

The primary customers of this product are Hospitals, Rheumatologists, and EMTs.

c. Patient-related concerns:

This algorithm must provide accurate diagnoses to prevent the consequences of a false negative or false positive result. Minimizing the number of inaccurate results is crucial as false negatives can lead to a patient not receiving the treatment that they need and false positives can lead to patients being exposed to unnecessary treatments and medications. It is also important that patient health information is not disclosed without proper notice as outlined in 45 CFR 164.520 [9].

d. Competition:

Other methods of detecting SjS include blood and urine tests, Schirmer tear test, Sialography, Salivary scintigraphy, and biopsy [10-14]. While these tests are less subjective than the current OMERACT grading system, they are significantly more invasive and time consuming than ultrasound scans. Additionally, a patent titled 'Method for Developing a Machine Learning Model of a Neural Network for Classifying Medical Images' by Tienovix LLC claims protection for a machine learning model relating to Data Collection, Feature Definition, Image Analysis, Labeling, Data Splitting, Neural Network Training, Training Metrics, Threshold Evaluation, Validation Process, Validation Metrics, and Model Storage [15]. This patent describes a method for obtaining medical image data, including ultrasound images, and trains a machine learning model to analyze features in the image and validate that model’s accuracy with a training set. This method can be applied to diagnose SjS by training a machine learning model to recognize features of salivary gland ultrasound scans and grade them based on their characteristics. Another patent titled “Machine-aided workflow in ultrasound imaging”, protects the use of computer-aided classification to detect objects inside of the body [16]. While this patent describes the classification of organs in an ultrasound scan, a similar model could be used to distinguish the salivary glands in ultrasound scans of potential SjS patients.

Reference

1. M. Ramos-Casals and J. Font, “Primary sjögren’s syndrome: Current and emergent AETIOPATHOGENIC concepts,” Rheumatology, vol. 44, no. 11, pp. 1354–1367, 2005. doi:10.1093/rheumatology/keh714

2. P. Brito-Zerón, S. Retamozo, and M. Ramos-Casals, “Síndrome de Sjögren,” Medicina Clínica, vol. 160, no. 4, pp. 163–171, 2023. doi:10.1016/j.medcli.2022.10.007

3. S. S. Kassan and H. M. Moutsopoulos, “Clinical manifestations and early diagnosis of Sjögren syndrome,” Archives of Internal Medicine, vol. 164, no. 12, p. 1275, 2004. doi:10.1001/archinte.164.12.1275

4. Lorenzon, Michele et al. “Salivary Gland Ultrasound in Primary Sjögren's Syndrome: Current and Future Perspectives.” Open access rheumatology : research and reviews vol. 14 147-160. 1 Sep. 2022, doi:10.2147/OARRR.S284763

5. 21 CFR § 56.102, 1981

6. O. of the Commissioner, “Informed consent for clinical trials,” U.S. Food and Drug Administration, https://www.fda.gov/patients/clinical-trials-what-patients-need-know/informed-consent-clinical-trials (accessed Sep. 21, 2023).

7. 45 CFR § 164.514, 2000

8. O. for C. R. (OCR), “Methods for de-identification of phi,” HHS.gov, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale (accessed Sep. 21, 2023).

9. 45 CFR § 164.520, 2000

10. “Blood and urine tests,” Johns Hopkins Sjögren’s Center, https://www.hopkinssjogrens.org/disease-information/diagnosis-sjogrens-syndrome/blood-and-urine-tests/ (accessed Sep. 19, 2023).

11. A.-L. Stefanski et al., “The diagnosis and treatment of Sjögren’s syndrome,” Deutsches Ärzteblatt international, 2017. doi:10.3238/arztebl.2017.0354

12. N. Ohbayashi, I. Yamada, N. Yoshino, and T. Sasaki, “Sjögren syndrome: Comparison of assessments with mr sialography and conventional sialography.,” Radiology, vol. 209, no. 3, pp. 683–688, 1998. doi:10.1148/radiology.209.3.9844659

13. I. Umehara, I. Yamada, Y. Murata, Y. Takahashi, N. Okada, and H. Shibuya, “Quantitative evaluation of salivary gland scintigraphy in Sjörgen’s syndrome,” Journal of Nuclear Medicine: Official Publication, Society of Nuclear Medicine, vol. 40, no. 1, pp. 64–69, Jan. 1999, Accessed: Sep. 22, 2023. [Online]. Available: https://www.google.com/url?q=https://pubmed.ncbi.nlm.nih.gov/9935059/&sa=D&source=docs&ust=1695404911950832&usg=AOvVaw1wYoNw1_pR1FXK9wno3T62

14. “Diagnosing sjogren’s syndrome,” Patient Care at NYU Langone Health, https://nyulangone.org/conditions/sjogrens-syndrome/diagnosis (accessed Sep. 19, 2023).

15. W. R. Buras, C. S. Russell, and K. Q. Nguyen, “Method for developing a machine learning model of a neural network for classifying medical images.” https://patents.google.com/patent/US11017695B2 (accessed Sep. 22, 2023).

16. 라오빔바 and 구라카이스마일 엠, “Machine-aided workflow in ultrasound imaging.” https://www.google.com/url?q=https://patents.google.com/patent/KR20190053807A/en&sa=D&source=docs&ust=1695404911953234&usg=AOvVaw0FAXbca4VpQdMv60yQYrQ2 (accessed Sep. 22, 2023).