Prediction for Periodontists by Oral Bacteria in Korean

2020 1st Semester Interdisciplinary Project

20161206 JaewoongLee

June 4, 2020

Student ID	20161206
Name	Jaewoong Lee
School	School of Electrical & Computer Engineering
1 Track	Computer Science & Engineering
2 Track	Bio-medical Engineering
Advisor 1	Prof. Sungahn Ko
Advisor 2	Prof. Semin Lee

Contents

List of Tables

List of Figures

1 Introduction

1.1 Periodontitis

Periodontitis is an inflammatory disease of the periodontium which is characterized by a progressive destruction of the tissues supporting the tooth [2]. In histopathologically, periodontitis may result periodontal pocketing, location of junctional epithelium apical to the cemento-enamal junction, loss of collagen fibers subjacent to the pocket epithelium, numerous poly-morphonuclear leukocytes in epithelium and a dense inflammatory cell infiltrate with plasma cells, lymphocytes, and macrophages [3]. Periodontits is currently assumed to progress as periodic, relatively short episodes of rapid tissue destruction followed by some prolonged intervening periods of disease remission [2].

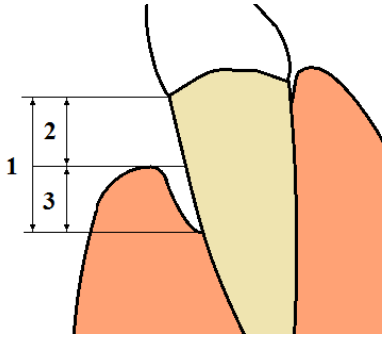


Figure 1: Diagram of Gingival Recession [1]

Periodontitis is diagnosed by measuring clinical attachment loss (AL). Note that the AL is the length of the figure 1-1, which is sum of gingival recession (GR) in figure 1-2, and probing depth (PD) in figure 1-3.

Periodontitis is generally believed to be a result of a host-parasite interaction in which bacteria are the determinants of periodontitis [4]. In etiology, the primary cause of periodontitis is presumed as a bacterial infection as the primary cause of periodontitis [3]. Thus, the treatment of periodontitis includes antibiotics and dental surgery.

In this manner, some medicines have been introduced for treatment. However, the success in the prevention and treatment of periodontitis has been limited. Many *in vitro* studies shows that Asian have the different bacteria from non-Asian, due to their groceries [5]. Thus, the developments of plaque and calculus in Asian differ, and may lead to distant reactions between Asian and non-Asian.

1.2 Machine Learning

Machine learning is the study of algorithms which advance spontaneously through experience. Machine learning is conjugated where is infeasible with conventional algorithms such as computer vision. Many papers show that machine learning brings out better result than human recognition.

If the feedback provides the correct answer for specific inputs, then learning problem is called supervised learning [6]. Classification is a kind of supervised learning for discrete values; regression is for continuous values.

1.3 Purpose of Research

There are many studies which have tried to find bacteria as bio-markers [7, 8]. Most of these papers, though, researched in Western people [9, 10]. As I mentioned herein-above, oral bacteria population may differ between Western and non-Western. In this approach, therefore, prediction periodontitis from machine learning which based on oral bacteria population of Korean is required.

I aimed to probe the performance of machine learning which predict the severity of periodontitis. Specifically, the purpose of this research is herein-after:

1. Classify the stage of periodontitis by oral bacteria.

2. Regress the AL, or the PD by oral bacteria.

2 Materials

2.1 Clinical Examinations

This study included 784 samples from who visited the Department of Periodontics, Pusan National University Dental Hospital, between August 2016 and March 2019. The study protocol was approved by the Institutional Review Board of Pusan National University Dental Hospital (PNUDH-2016-019). All samples are provided written informed consent upon complete information regarding the objectives and procedures of this study.

The diagnosis of samples was completed as [11]. Also, the stage of periodontitis was categorized on the basis of the AL as following:

- Healthy: $\leq$ 1mm
- Slight: 1-2 mm
- Moderate: 3-4 mm
- Severe: $\geq$ 5 mm

Moreover, the following patients were excluded:

- who received periodontal treatment with past six months
- who were pregnant or breastfeeding
- who refused to approve the informed consent form

The AL was measured with a periodontal probe (PGF-W, Osung, Kwangmyung, Republic of Korea) during the clinical evaluation. All measurement were performed by two fully-experienced periodontitsts.

2.2 Analysis of Bacterial Copy

Collection of mouthwash sample and DAN extraction were performed as [12]. Also, the nine pathogens were chosen as herein-after:

1. *Porphyromonas gingivalis (Pg)*
2. *Tannerella forsysthia (Tf)*
3. *Treponema denticola (Td)*
4. *Prevotella intermedia (Pi)*
5. *Fusobacterium nucleatum (Fn)*
6. *Campylobacter rectus (Cr)*
7. *Aggregatibacter actinomycetemcomitans (Aa)*
8. *Peptostreptococcus anaerobius (Pa)*
9. *Eikenella corrodens (Ec)*

Multiplex qPCR system was optimized for the nine pathogens after the building of standard curves for each pathogen.

3 Methods

The entire program is disclosed by GitHub in `https://github.com/Fumire/Periodontist_Fall2019`.

3.1 Python Packages

Python programming language had been used to analyze data. Also, many Python modules had been adopted as hereinafter.

3.1.1 Pandas

Pandas is a Python library of rich data structures and tools for working with structured data sets common to statistics, finances, social sciences, and many other fields [13].

3.1.2 Scikit-learn: Machine Learning in Python

Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems [14].

3.1.3 Seaborn

Seaborn is a Python data visualization library based on *matplotlib*. It provides a high-level interface for drawing attractive and informative statistics graphics [15].

3.2 Classification

In classification, every combination of the nine pathogens, $2^9 = 512$ combinations, will be used. Also, in classification, some classes are merged into new class. For instance, healthy class and slight class could be merged, then the algorithm will be performed with four classes. As the pathogen combination, every combination of merging classes will be finished.

3.2.1 Confusion Matrix and Its Derivations

A confusion matrix is a table which displays the performance of classification algorithm. Typically, the confusion matrix is like as table 1.

Table 1: Abstract Form of Confusion Matrix

		Actual Class	
		Positive	Negative
Predicted Class	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Many derivations, such as sensitivity and specificity, come from the confusion matrix. The equation of derivations are followings:

- Sensitivity = $\frac{TP}{P}$
- Specificity = $\frac{TN}{N}$
- Precision = $\frac{TP}{TP+FP}$
- Negative predictive value = $\frac{TN}{TN+FN}$
- Miss rate = $\frac{FN}{P} = \frac{FN}{FN+TP}$
- False positive rate = $\frac{FP}{N} = \frac{FP}{FP+TN}$
- False discovery rate = $\frac{FP}{FP+TP}$
- False omission rate = $\frac{FN}{FN+TN}$
- Threat score = $\frac{TP}{TP+FN+FP}$

- Accuracy = $\frac{TP+TN}{P+N}$
- F1 score = $\frac{2TP}{2TP+FP+FN}$

Note that followed abbreviations are used:

- P: Positive
- N: Negative
- TP: True Positive
- TN: True Negative
- FP: False Positive
- FN: False Negative

3.2.2 Classification Algorithm

For classification, the followed algorithms has been used:

- K-Neighbors
- Linear Support Vector Classification (SVC)
- Poly SVC
- RBF SVC
- Sigmoid SVC
- Decision Tree
- Random Forest
- Adam Neural Network (NN)
- lbfgs NN
- Ada-Boost

These are almost every algorithm which are supported by *Scikit-learn*.

3.3 Regression

In regression, every combination of the nine pathogens, $2^9 = 512$ combinations, will be used.

3.3.1 Welch's t-test

t-test is a two-sample test which is used to test a hypothesis that two populations have identical means. Student's t-test is commonly used. Though, Welch's t-test is more robust then Student's t-test [16]. Therefore, Welch's t-test is adopted to verify null hypothesis.

3.3.2 Coefficient of Determination

Coefficient of determination, also known as R^2 or R-square, is common to use as an index of the size of the relation [17].

3.3.3 Regression Algorithm

For regression, the followed algorithm has been used:

- Linear Regression
- Ridge
- Support Vector Regression (SVR)
- Nu SVR
- Linear SVR
- Elastic Network
- K-Neighbors
- Decision Tree
- lbfgs Multi-Layer Perceptron (MLP)
- sgd MLP

These are almost every algorithm which are supported by *Scikit-learn*.

4 Results

4.1 5-class Classification

Figure 2 displays the derivations of confusion matrix in 5-class classification. Note that the values in figure 2, mean values from combinations which used same number of features will be shown.

As figure 2, the Poly-SVC algorithm has the best values with all features.

Figure 3 shows the heatmap between real and predicted classes. As shown as figure 2, the accuracy is over 80 %, and wrong predicted class is predicted neighbor class; for instance, healthy and moderate class are neighbor class of slight class. Moreover, acute class is commonly classified as moderate and severe class.

4.2 4-class Classification

4.2.1 Merged Healthy-Slight Class

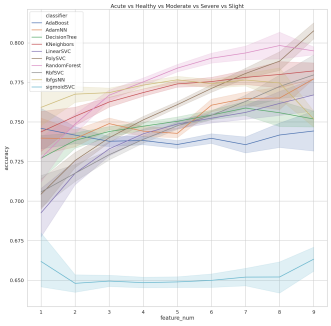
Figure 4 displays the derivations of confusion matrix in merged healthy-slight classification. Note that the values in figure 4, mean values from combination which used same number of features will be shown. As figure 4, the Random Forest algorithm has the best values with all features. Thus, the heatmap plot with real and predicted classes is shown as figure 5.

4.2.2 Merged Slight-Moderate Class

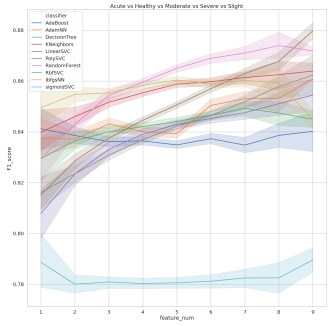
Figure 6 displays the derivations of confusion matrix in merged slight-moderate classification. Note that the values in figure 6, mean values from combination which used same number of features will be shown. As figure 6, the K-Neighbor algorithm has the best values with all features. Thus, the heatmap plot with real and predicted classes is shown as figure 7.

4.2.3 Merged Moderate-Severe Class

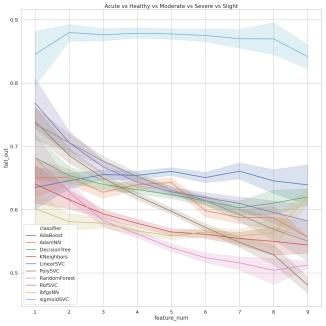
Figure 8 displays the derivations of confusion matrix in merged moderate-severe classification. Note that the values in figure 8, mean values from combination which used same number of features will be shown. As figure 8, the Random Forest algorithm has the best values with all features. Thus, the heatmap plot with real and predicted classes is shown as figure 9.



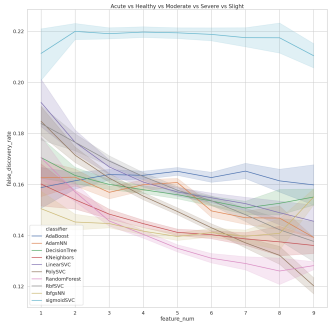
Accuracy



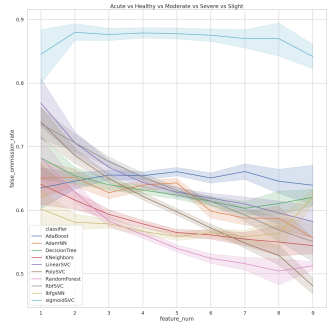
F1 score



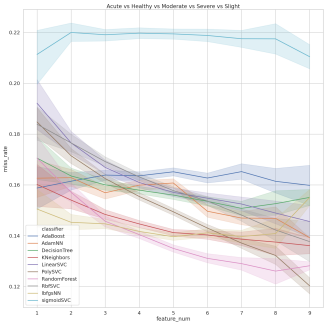
Fall-out



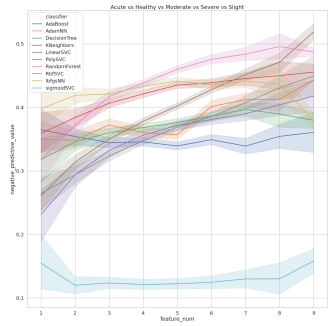
False discovery rate



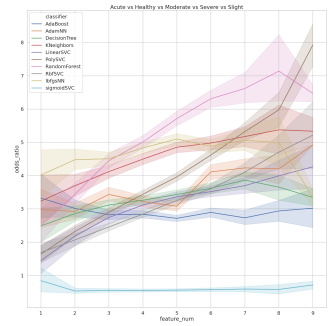
False ommision rate



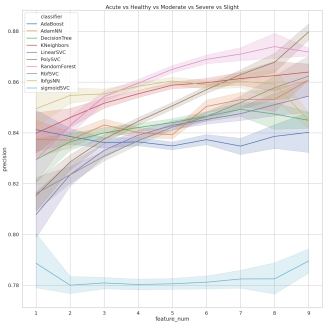
Miss rate



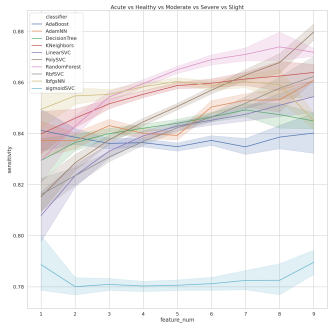
Negative predictive value



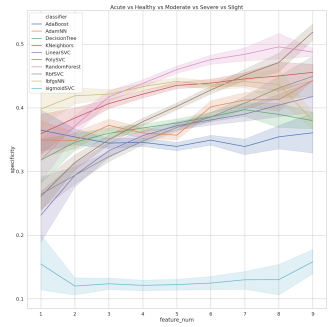
Odds ratio



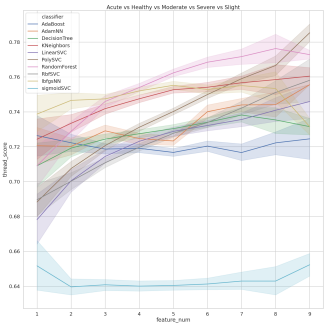
Precision



Sensitivity



Specificity



Thread score

Figure 2: Confusion Matrix Derivations from 5-class Classification

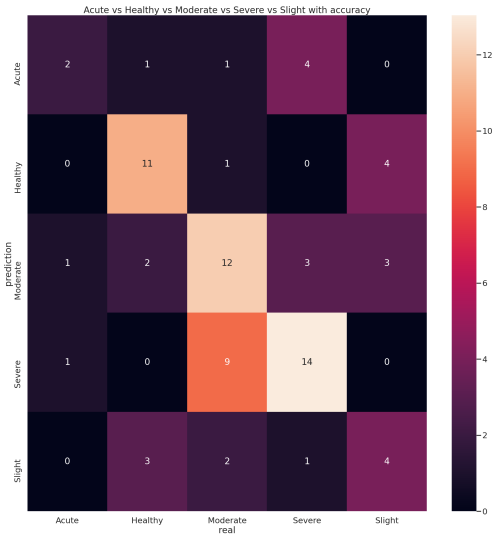


Figure 3: Heatmap Plot for 5-class Classification with Poly-SVC

4.2.4 Merged Slight-Acute Class

Figure 10 displays the derivations of confusion matrix in merged slight-acute classification. Note that the values in figure 10, mean values from combination which used same number of features will be shown. As figure 11, the Random Forest algorithm as the best values with all features. Thus, the heatmap plot with real and predicted classes is shown as figure 11.

4.2.5 Merged Moderate-Acute Class

Figure 12 displays the derivations of confusion matrix in merged moderate-acute classification. Note that the values in figure 12, mean values from combination which used same number of features will be shown. As figure 12, the Random Forest algorithm has the best values with all features. Thus, the heatmap plot with real and predicted classes is shown as figure 13.

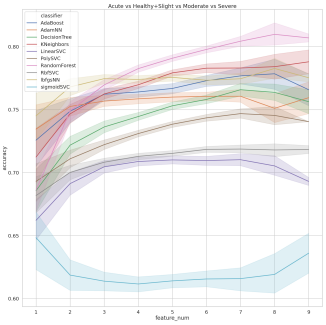
4.2.6 Merged Severe-Acute Class

Figure 14 displays the derivations of confusion matrix in merged severe-acute classification. Note that the values in figure 14, mean values from combination which used same number of features will be shown. As figure 14, the RBF-SVC algorithm has the best values with all features. Thus, the heatmap plot with real and predicted classes is shows as figure 15.

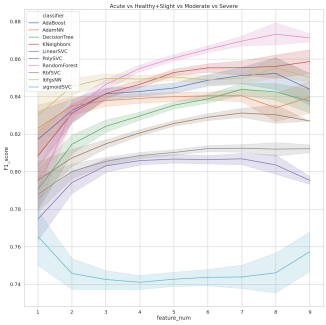
4.3 Distribution of Depth

In regression, the AL and the PD will be used. Thus, it is required to inspect about the AL and the PD. Figure 16 shows the scatter map between the AL and the PD. Also, figure 17 displays the violin plot between depth and class. Note that follow annotations for figure 17:

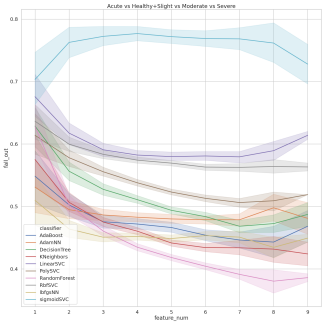
- *: $1.00 \times 10^{-2} < p \leq 5.00 \times 10^{-2}$
- **: $1.00 \times 10^{-3} < p \leq 5.00 \times 10^{-2}$
- ***: $1.00 \times 10^{-4} < p \leq 5.00 \times 10^{-3}$
- ****: $p \leq 5.00 \times 10^{-4}$



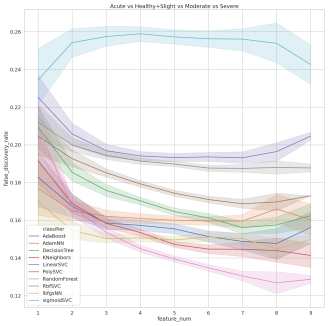
Accuracy



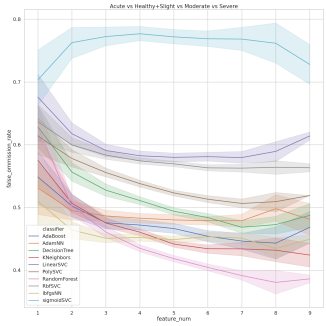
F1 score



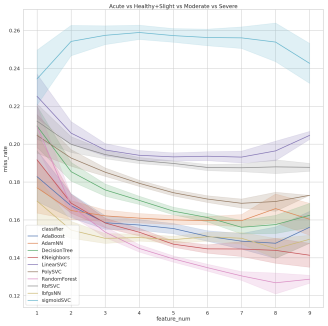
Fall-out



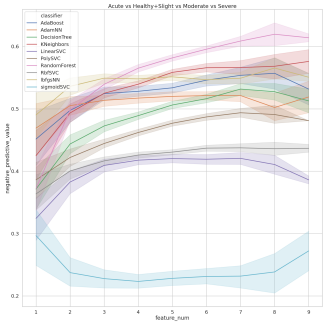
False discovery rate



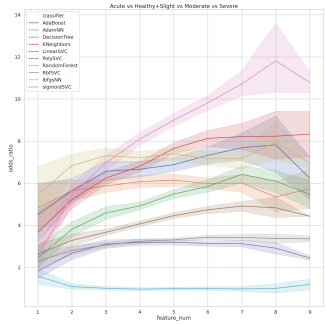
False ommision rate



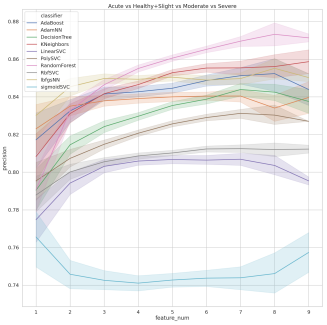
Miss rate



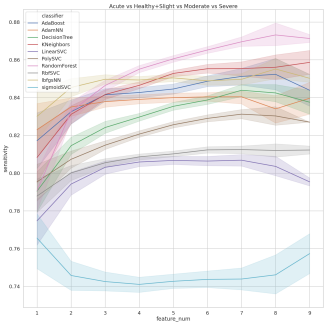
Negative predictive value



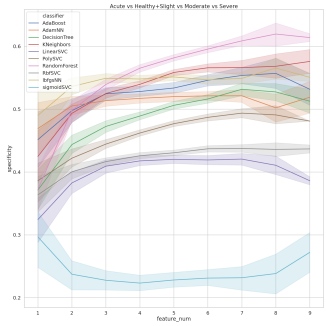
Odds ratio



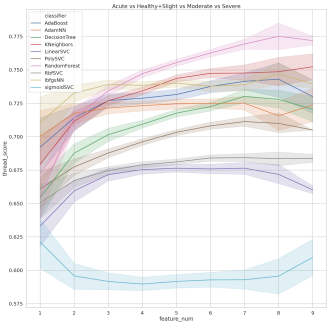
Precision



Sensitivity



Specificity



Thread score

Figure 4: Confusion Matrix Derivations from Merged Healthy-Slight Classification

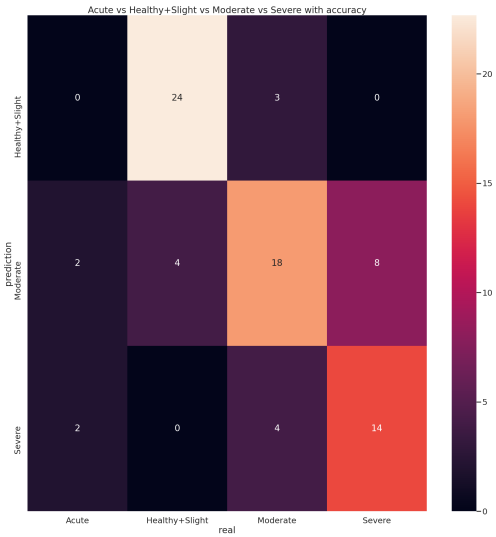


Figure 5: Heatmap Plot for Merged Healthy-Slight Classification with Random Forest

4.4 Regression

Figure 18 displays the R-square value in regression.

5 Discussion

5.1 Classification

Many intense studies [18, 19] imply that there are statistically differences of bacteria population between each stage of periododontitis. There are no doubt in which machine learning process will find the strategy to classify amongst periodontal classes. It is sustained as figure 2, the accuracy is over 80 %. Moreover, as figure 3, most of missed prediction has been predicted neighbor classes. Thus, it is believed that many indicators such as accuracy will be better while combining neighbor classes.

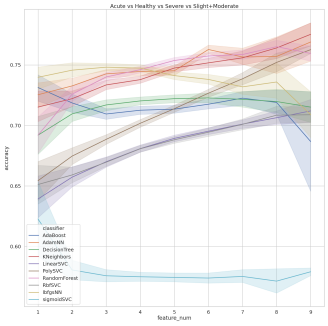
Figures 4, 6 and 8 show that the indicators are far from improving. Only three classes are predicted as figures 4, 6 and reffig:m-s-heatmap; in other words, the acute class was dropped from prediction. This is happened by two major reasons. First, there are only four cases in acute class. Second, as figure 2, half of all cases from acute class have similar bacteria population with moderate or severe class. Therefore, classification algorithms get high accuracy even though dropping acute class.

Moreover, as mentioned herein-above, acute class has analogous bacteria population with moderate or severe class. Thereupon, it is expected that enhanced result from merged moderate-acute and severe-acute class. As figures 11, 13 and 15, every four classes were predicted; there are no dropped class. Merging severe-acute class has better accuracy than merging moderate-acute class for comparing figure 12 and figure 14.

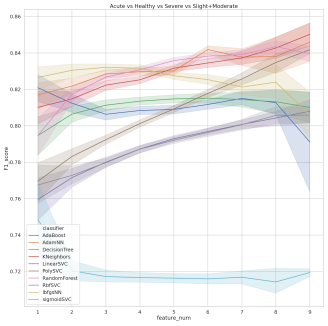
The number of used features is main factor for classification. There is a trend which indicator is being better along as many used features. Thus, it is required to perform the process with not only the nine bacteria more bacteria such as *Prevotell intermedia* [20] which is related with periodontitis.

5.2 Regression

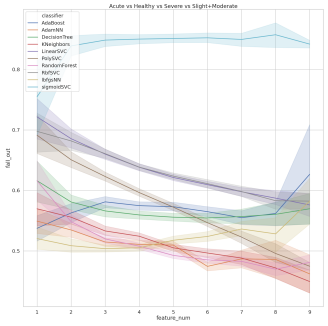
In figure 16, there is a linear trend on the AL and the PD. Many subjects have same value of the AL and the PD; according to figure 1, it implies that many subjects has no GR in their tooth. Also, there are statistically significant differences between each class as figure 17. Thus, regression process is expected good result as classification process does.



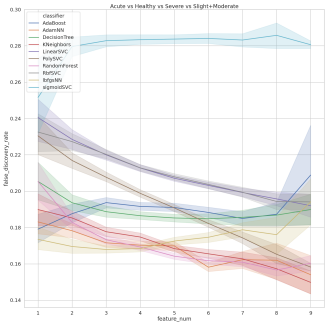
Accuracy



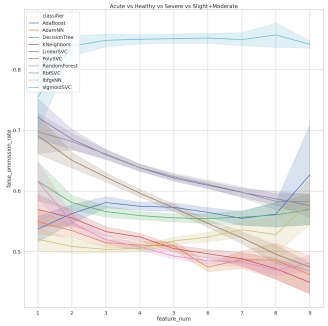
F1 score



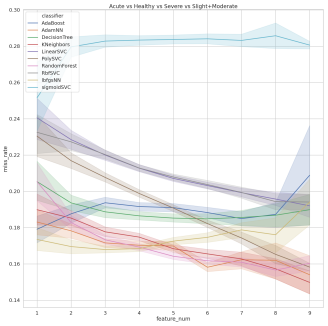
Fall-out



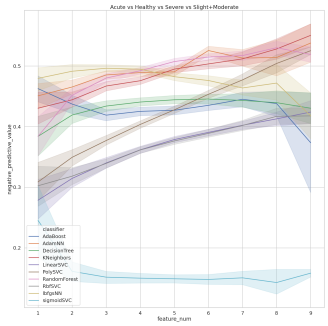
False discovery rate



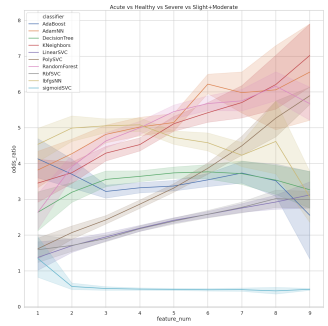
False ommision rate



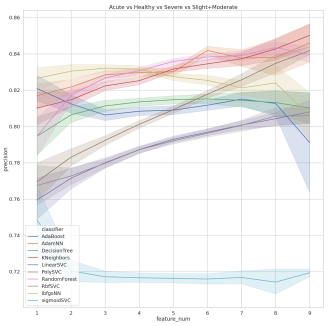
Miss rate



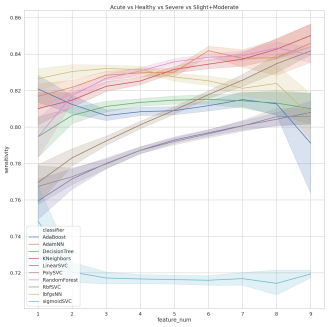
Negative predictive value



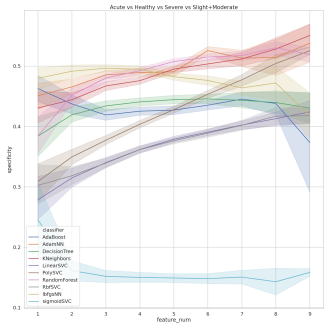
Odds ratio



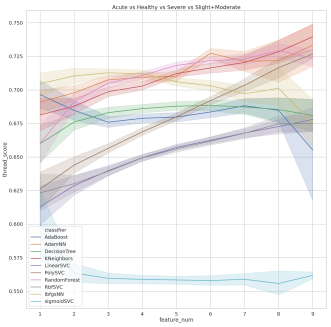
Precision



Sensitivity



Specificity



Thread score

Figure 6: Confusion Matrix Derivations from Slight-Moderate Classification

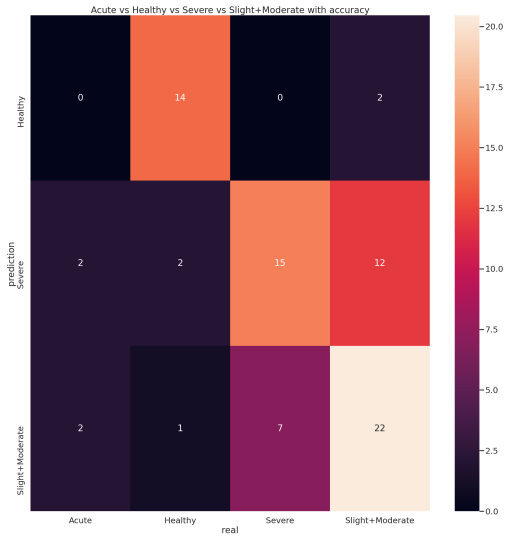


Figure 7: Heatmap Plot for Merged Slight-Moderate Classification with K-Neighbor

However, the result is different as figure 18. The R-square values are near 0.5; this implies that the weak relationship between bacteria and the AL or the PD. Also, the relationship between the number of features and R-square value is not strong as classification. It means that there are driver bacteria, which have major roles in periodontitis; and, the AL and the PD are effected by these driver bacteria.

6 Acknowledgment

The relative study which based on the identical data has been submitted *American Society for Microbiology* as "Prediction of chronic periodontitis severity using machine learning models based on salivary bacterial copy number".

I thank all study subjects for their generous participation and the clinicians for their contributions leading to the successful completion of this study. This work was partly supported by the Technological Innovation R&D Program (C0445482), funded by the Small and Medium business Administration (SMBA, Republic of Korea). This work also partly supported by the Next-Generation Information Computing Development Program of the National Research Foundation of Korea funded by the Ministry of Science and ICT (NRF-2016M3C4A7952635). This research work was also partly supported by the National Research Foundation (NRF) of Korea grant NRF-2017M3A9B6062026, funded by the government of Republic of Korea. I would like to thank David Whee-Young Choi for constructive criticism of the manuscript.

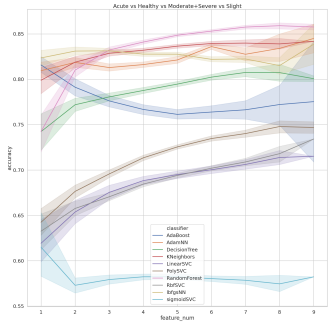
References

[1] "Periodontal terms diagram gingival recession," Mar 2014. [Online]. Available: https://en.wikipedia.org/wiki/File:Periodontal_terms_diagram_gingival_recession.png

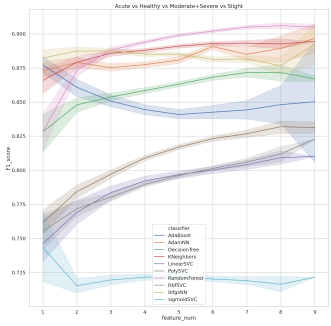
[2] M. A. Listgarten, "Pathogenesis of periodontitis," *Journal of clinical periodontology*, vol. 13, no. 5, pp. 418–425, 1986.

[3] T. F. Flemmig, "Periodontitis," *Annals of Periodontology*, vol. 4, no. 1, pp. 32–37, 1999.

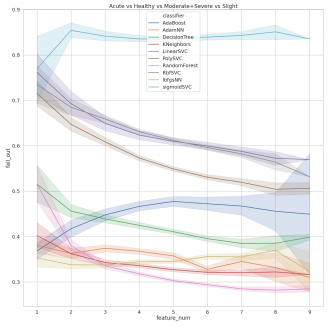
[4] N. G. Clarke and R. S. Hirsch, "Personal risk factors for generalized periodontitis," *Journal of clinical periodontology*, vol. 22, no. 2, pp. 136–145, 1995.



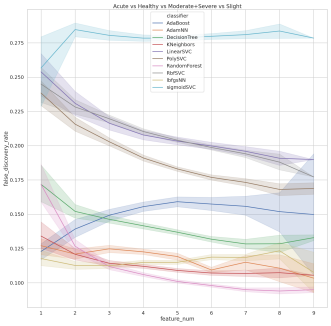
Accuracy



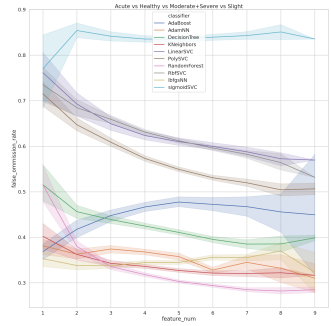
F1 score



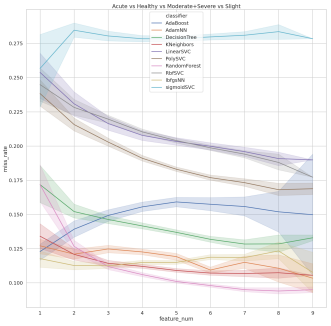
Fall-out



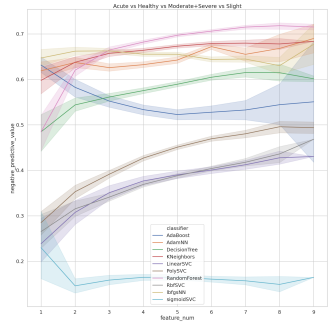
False discovery rate



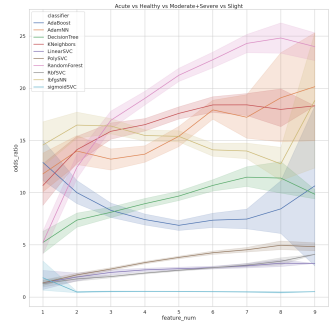
False ommision rate



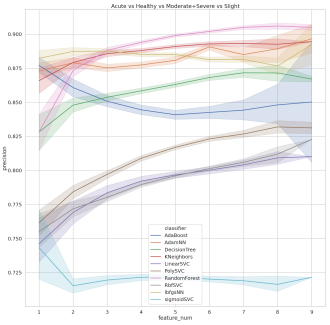
Miss rate



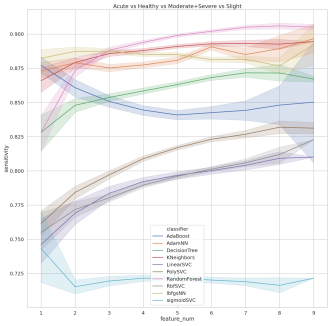
Negative predictive value



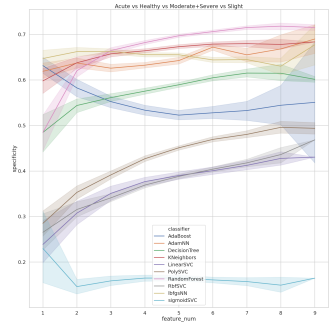
Odds ratio



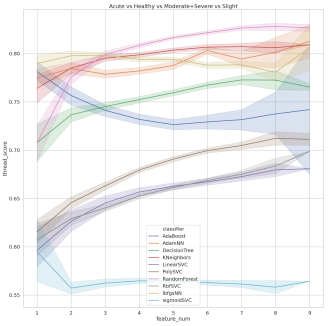
Precision



Sensitivity



Specificity



Thread score

Figure 8: Confusion Matrix Derivations from Moderate-Severe Classification

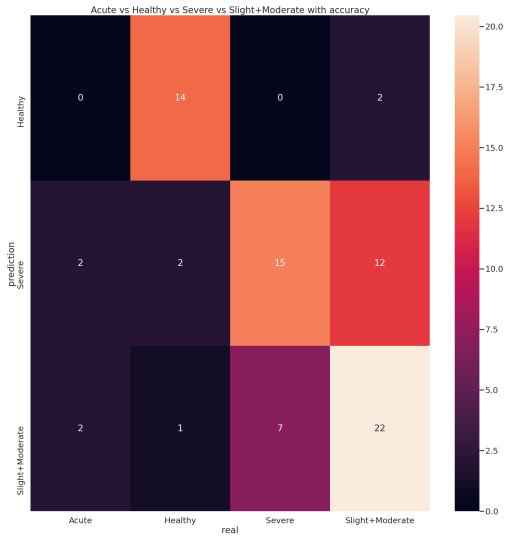


Figure 9: Heatmap Plot for Merged Moderate-Severe Classification with Random Forest

[5] A. T. Borchers, T. K. Mao, C. L. KEEN, H. H. SCHMITZ, H. WATANABE, and M. E. GERSHWIN, "Traditional asian medicine and oral health," *Journal of Traditional Medicines*, vol. 21, no. 1, pp. 17–26, 2004.

[6] S. Russell and P. Norvig, "Artificial intelligence: A modern approach prentice-hall," *Englewood cliffs, NJ*, vol. 26, 1995.

[7] L. Wolff, G. Dahlén, and D. Aeppli, "Bacteria as risk markers for periodontitis," *Journal of periodontology*, vol. 65, pp. 498–510, 1994.

[8] A. C. R. Tanner, C. Haffer, G. Bratthall, R. Visconti, and S. Socransky, "A study of the bacteria associated with advancing periodontitis in man," *Journal of clinical periodontology*, vol. 6, no. 5, pp. 278–307, 1979.

[9] C. S. Miller, C. P. King Jr, M. C. Langub, R. J. Kryscio, and M. V. Thomas, "Salivary biomarkers of existing periodontal disease: a cross-sectional study," *The Journal of the American Dental Association*, vol. 137, no. 3, pp. 322–329, 2006.

[10] M. Taba, J. Kinney, A. S. Kim, and W. V. Giannobile, "Diagnostic biomarkers for oral and periodontal diseases," *Dental Clinics*, vol. 49, no. 3, pp. 551–571, 2005.

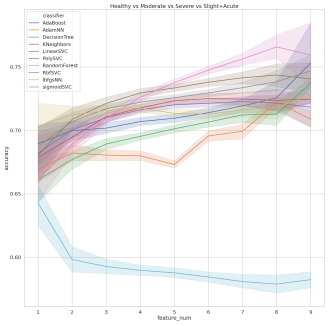
[11] G. C. Armitage, "Development of a classification system for periodontal diseases and conditions," *Annals of periodontology*, vol. 4, no. 1, pp. 1–6, 1999.

[12] E.-H. Kim, J.-Y. Joo, Y. J. Lee, J.-K. Koh, J.-H. Choi, Y. Shin, J. Cho, E. Park, J. Kang, K. Lee *et al.*, "Grading system for periodontitis by analyzing levels of periodontal pathogens in saliva," *PloS one*, vol. 13, no. 11, 2018.

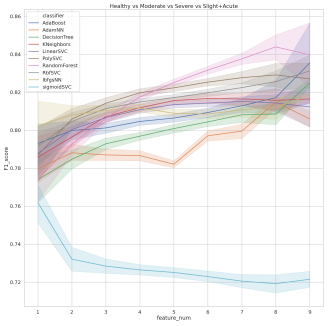
[13] W. McKinney *et al.*, "pandas: a foundational python library for data analysis and statistics," *Python for High Performance and Scientific Computing*, vol. 14, no. 9, 2011.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

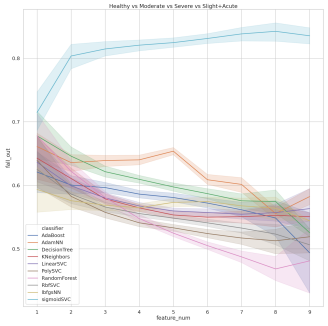
[15] M. Waskom, O. Botvinnik, P. Hobson, J. B. Cole, Y. Halchenko, S. Hoyer, A. Miles, T. Augspurger, T. Yarkoni, T. Megies, L. P. Coelho, D. Wehner, cynddl, E. Ziegler, diego0020, Y. V. Zaytsev, T. Hoppe,



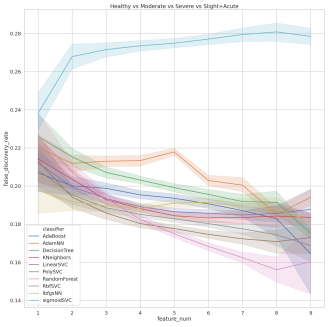
Accuracy



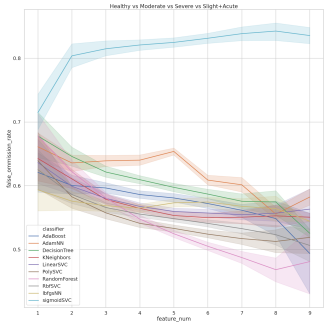
F1 score



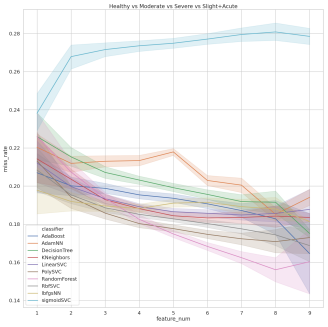
Fall-out



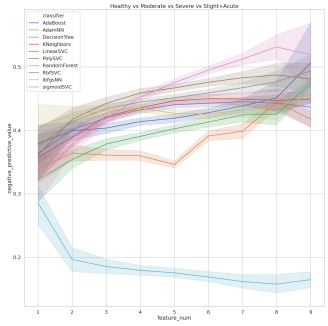
False discovery rate



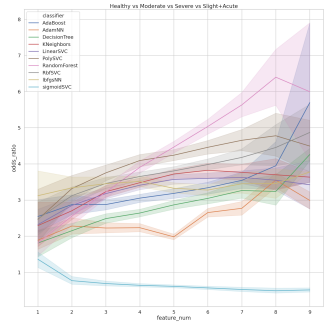
False ommision rate



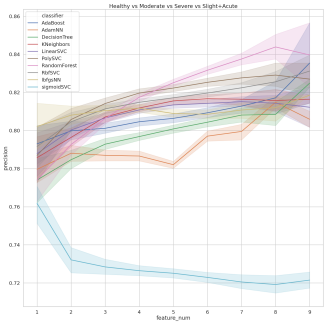
Miss rate



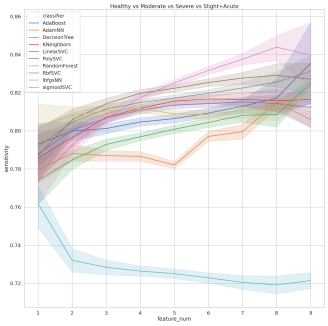
Negative predictive value



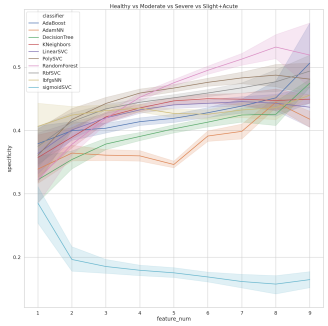
Odds ratio



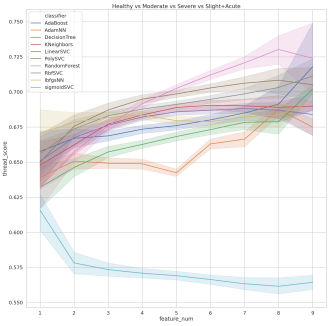
Precision



Sensitivity



Specificity



Thread score

Figure 10: Confusion Matrix Derivations from Slight-Acute Classification

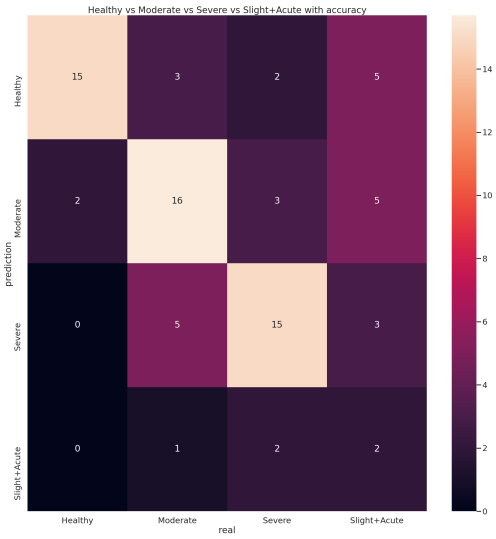


Figure 11: Heatmap Plot for Merged Slight-Acute Classification with Random Forest

S. Seabold, P. Cloud, M. Koskinen, K. Meyer, A. Qalieh, and D. Allan, "seaborn: v0.5.0 (november 2014)," Nov. 2014. [Online]. Available: https://doi.org/10.5281/zenodo.12710

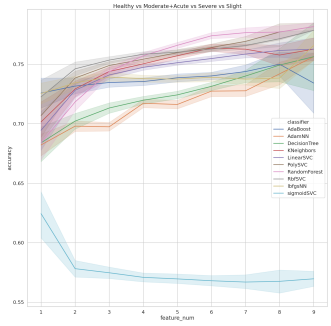
[16] M. Delacre, D. Lakens, and C. Leys, "Why psychologists should by default use welch's t-test instead of student's t-test," *International Review of Social Psychology*, vol. 30, no. 1, 2017.

[17] D. J. Ozer, "Correlation and the coefficient of determination." *Psychological bulletin*, vol. 97, no. 2, p. 307, 1985.

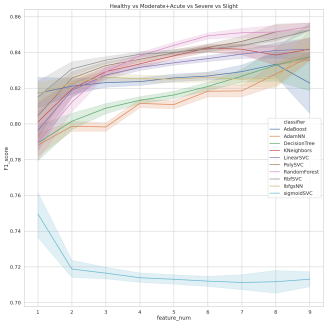
[18] A. C. Tanner, R. Kent Jr, E. Kanasi, S. C. Lu, B. J. Paster, S. T. Sonis, L. A. Murray, and T. E. Van Dyke, "Clinical characteristics and microbiota of progressing slight chronic periodontitis in adults," *Journal of clinical periodontology*, vol. 34, no. 11, pp. 917–930, 2007.

[19] B. Signat, C. Roques, P. Poulet, and D. Duffaut, "Role of fusobacterium nucleatum in periodontal health and disease," *Curr Issues Mol Biol*, vol. 13, no. 2, pp. 25–36, 2011.

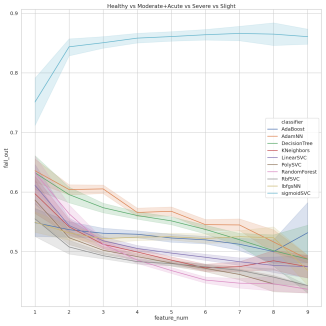
[20] N. J. López, "Occurrence of actinobacillus actinomycetemcomitans, porphyromonas gingivalis, and prevotella intermedia in progressive adult periodontitis," *Journal of periodontology*, vol. 71, no. 6, pp. 948–954, 2000.



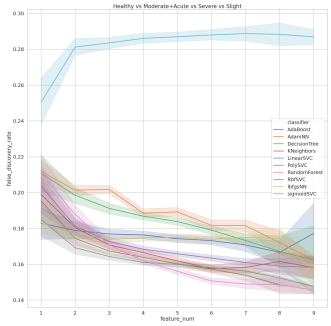
Accuracy



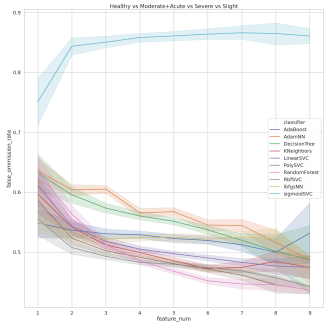
F1 score



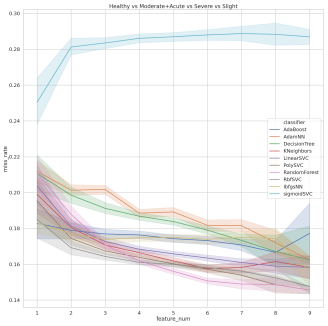
Fall-out



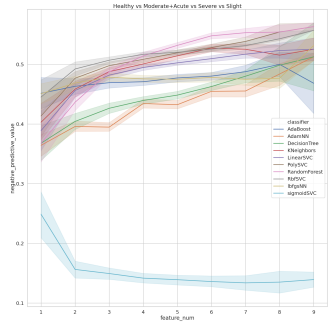
False discovery rate



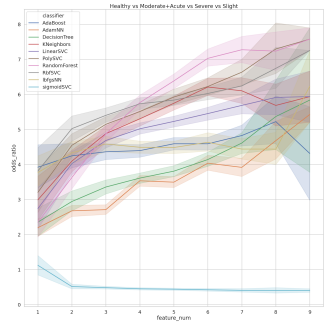
False ommision rate



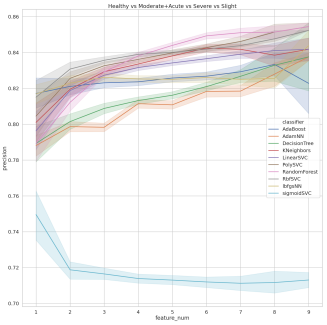
Miss rate



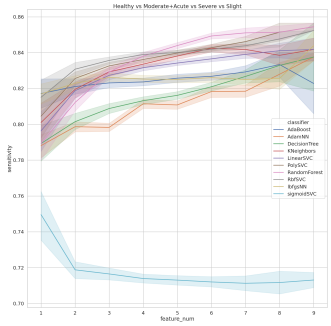
Negative predictive value



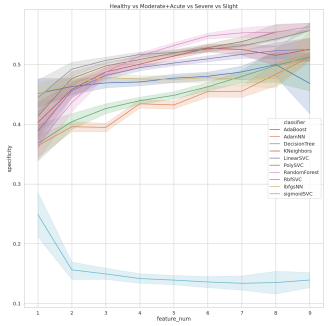
Odds ratio



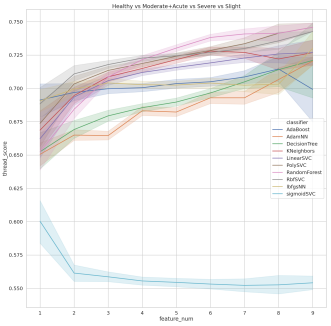
Precision



Sensitivity



Specificity



Thread score

Figure 12: Confusion Matrix Derivations from Moderate-Acute Classification

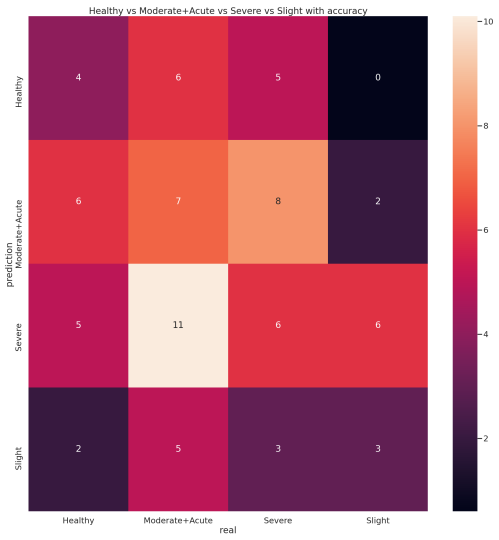
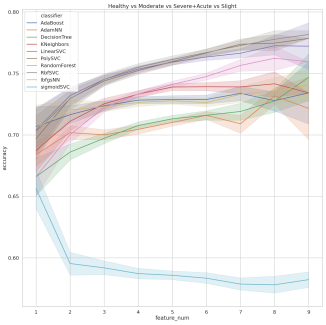
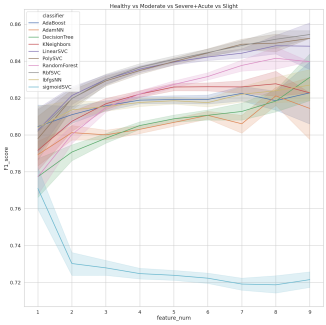


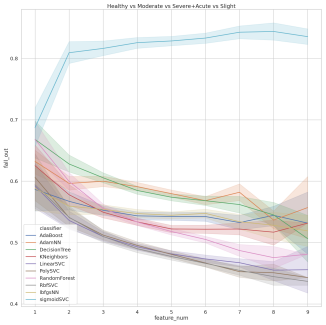
Figure 13: Heatmap Plot for Merged Moderate-Acute Classification with Random Forest



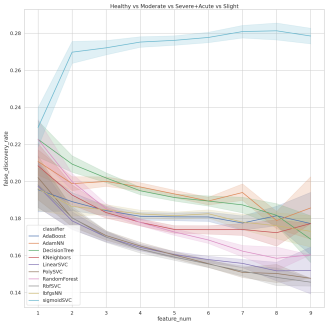
Accuracy



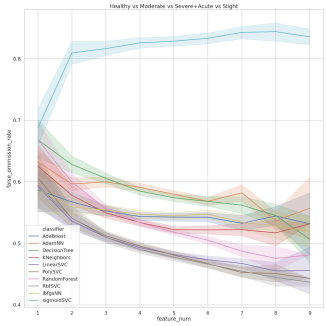
F1 score



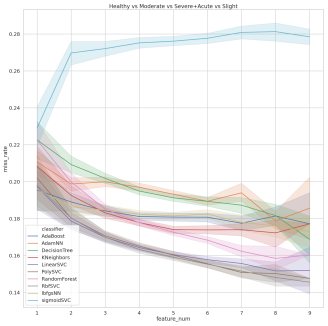
Fall-out



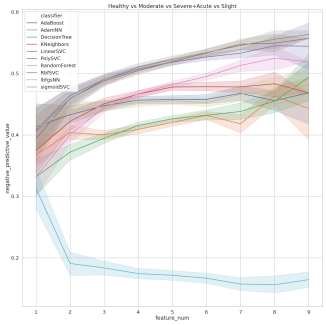
False discovery rate



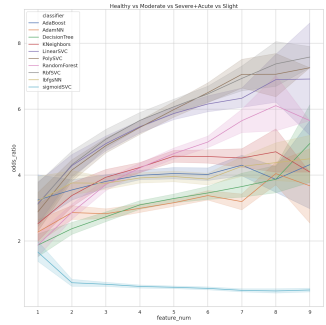
False ommision rate



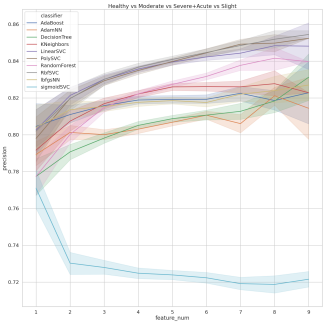
Miss rate



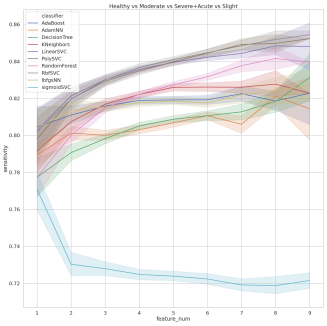
Negative predictive value



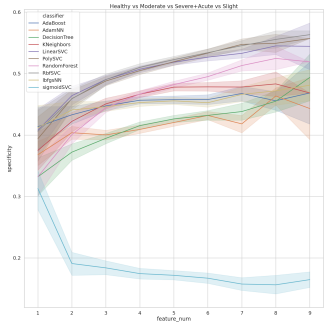
Odds ratio



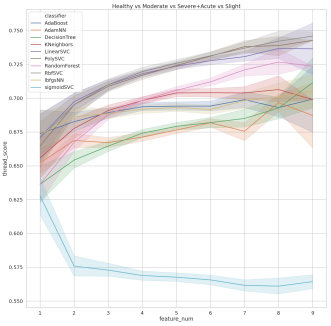
Precision



Sensitivity



Specificity



Thread score

Figure 14: Confusion Matrix Derivations from Severe-Acute Classification

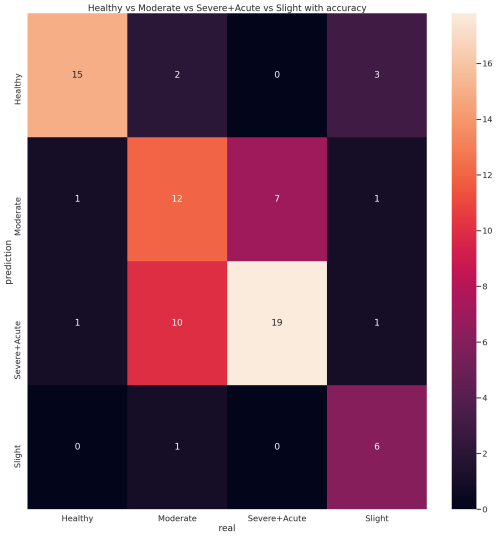


Figure 15: Heatmap Plot for Merged Severe-Acute Classification with RBF-SVC

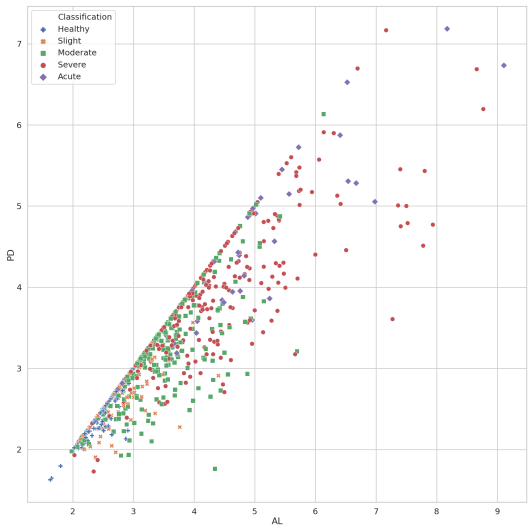
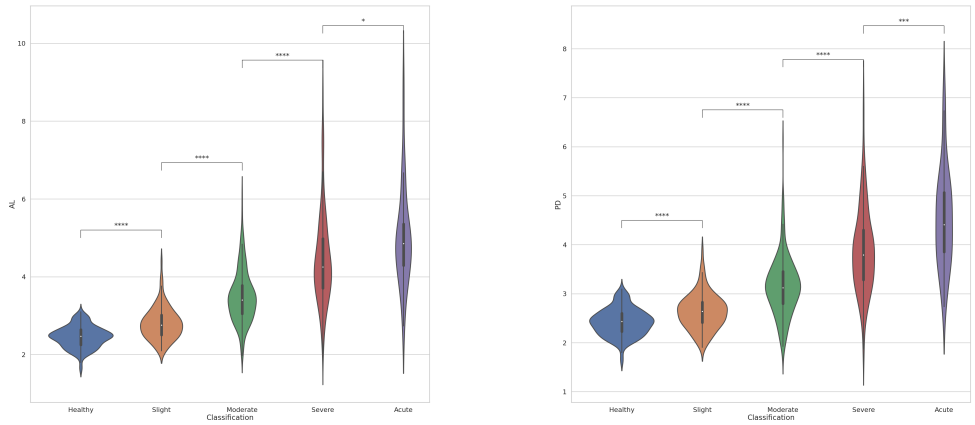
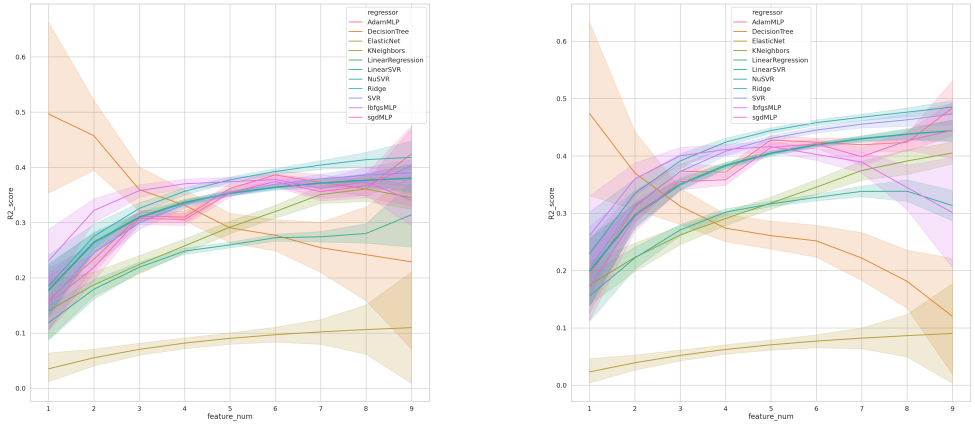


Figure 16: Scatter Map between AL and PD



(a) AL (b) PD

Figure 17: Violin Plot between Depth and Class



(a) AL (b) PD

Figure 18: R-Square Value of Regression