Classification of Lymphographic Data with Hybrid Models of Specialized Classifiers

Shubham Pateria
G1701615L
School of Computer Science and Engineering
Nanyang Technological University

I. INTRODUCTION

This report presents an analysis of the Lymphography data. The data was obtained from the *University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia*. Data is from Oncology domain which concerns with various forms of cancer and their treatment. In this particular case, the pathology of interest is Lymphoma or lymph-node cancer. The date provides various samples of observations of lymph-nodes and corresponding categories expressing if the lymph is *normal* or *affected.* A pathological lymph may indicate two different possibilities. A lymph may be *malignant.* In this case, the cancer starts at the lymph nodes and spreads out. In other cases, cancer may start at other sites in the body and spread into lymph nodes, in which case the pathological lymph site is called *metastases*. Given medical observations, correct identification of these different states is crucial for precise diagnosis.

This report discusses classification models built to categorize different pathological states of lymph. The goal of designed models is to correctly classify various lymph states. The data has some interesting characteristics such as discreteness and high imbalance of class distribution. Hybrid architectures are proposed to perform one-vs-all classification in different stages. At each stage, optimum classifier for a specific category is used. One of the architectures is a hybrid of ANFIS models. Instead of using generic ANFIS, Subtractive Clustering is employed to better handle the discrete structure of the input-space.

In later sections, different architectures are compared based their prediction performance.

II. ANALYSIS AND ENCODING OF DATA

The Lymphoma data consists of 148 instances with 18 attributes each. The attributes are of either nominal type or ordinal type. The full list of attributes is as follows:

No.	Attribute	Values	Type
1	lymphatics	normal, arched, deformed, displaced	nominal
2	block of affere	no, yes	nominal
3	bl. of lymph. c	no, yes	nominal
4	bl. of lymph. s	no, yes	nominal
5	by pass	no, yes	nominal
6	extravastases	no, yes	nominal
7	regeneration of	no, yes	nominal
8	early uptake in	no, yes	nominal
9	lym.nodes dimin	0-3	ordinal
10	lym.nodes enlar	1-4	ordinal
11	changes in lym.	bean, oval, round	nominal
12	defect in node	no, lacunar, lac. marginal, lac. central	nominal
13	changes in node	no, lacunar, lac. margin, lac. central	nominal
14	changes in stru	no, grainy, drop-like, coarse, diluted, reticular, stripped, faint	nominal
15	special forms	no, chalices, vesicles	nominal
16	dislocation of	no, yes	nominal
17	exclusion of no	no, yes	nominal
18	no. of nodes in	0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, >=70	ordinal

Table 1. Profile of data features.

The 148 data samples are divided into four classes with class distribution as:

Class:	Number of Instances:
normal find:	2
metastases:	81
malign lymph:	61
fibrosis:	4

Apparently, the dataset is imbalanced. Given that the classification problem concerns Lymphatic cancer, it is highly desirable that the normal class is correctly identified to avoid False Alarms. However, two data points are certainly not enough to learn a model capable of generalization. Hence, we need to look for sufficiently invariant and unique attributes.

After analyzing the attribute-wise data distributions, it was found that feature 1: ‘lymphatics’ has following distribution,

	normal	arched	deformed	displaced
normal find	**2**	0	0	0
metastases	0	39	26	16
malign lymph	0	28	16	17
fibrosis	0	0	4	0

lymphatics is the only attribute that was found to be uniquely capable of identifying *normal find* class. This may or may not be true for a broader dataset, however for the concerned data, this attribute is sufficient for a deterministic classification. Hence, for all models, ***normal find* class (class 1) is identified using a simple check of the *lymphatics* attribute.**

A. Encoding

The *nominal* variables in the data are encoded into one-hot vectors. The three ordinal variables (9, 10, and 18) are encoded as the integers provided in the data. One-hot or binary encoding is not done for ordinal variables in order to preserve the distance/magnitude order of the data values. **Post encoding the dimensionality of the feature space increases to 47.**

III. LEARNING MODELS

This models used for this study are *hybrid architectures*. As mentioned previously, the first check is for class 1: *normal find.* This check is a simple IF-ELSE rule based on the *lymphatics* attribute. Then, second-stage classifier is used to categorize class 4: *fibrosis* against the remaining classes – *metastases and malign*. If the data is identified as not belonging to *class 4,* then it is passed to a stage-3 classifier which classifies between *metastases* and *malign.* The two-stage architecture is required because a full 3-class classification fails to provide the best performance results.

*A. **MODEL 1 : RUSBoost and non-connectionist classifiers***

This model is shown in Fig. 1. The stage-2 classifier is a RUSBoost model [1]. This is a hybrid sampling/boosting algorithm specialized for handling imbalanced classes. The fundamental algorithm is an extension of *Random Undersampling*. The model is chosen because significant class imbalance is present in the given data, with *fibrosis* class having only 4 examples. The stage-3 classifier is a simple two-class classification system which can benefit from various algorithms such as (i) k-Nearest Neighbors with Cosine similarity (*because data is categorical*), or (ii) non-linear SVM.

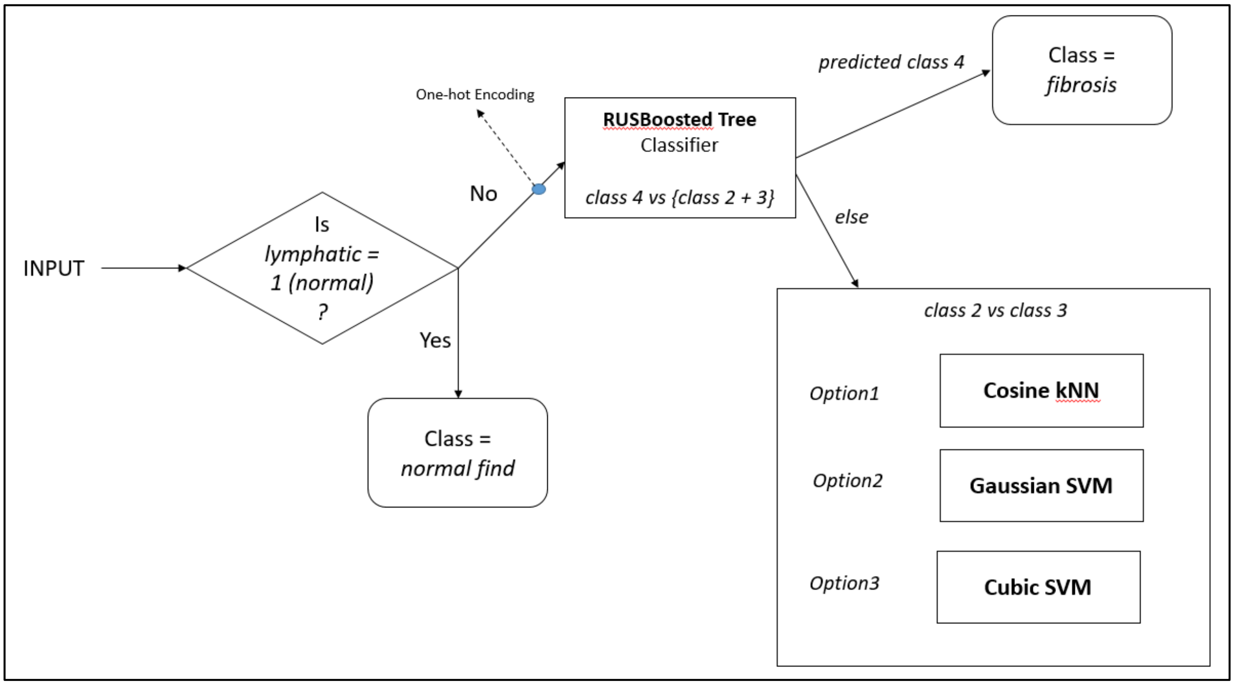


Fig. 1 Model - 1 Structure

B. *MODEL 2: ANFIS*

ANFIS [2] is a popular class of *Neuro Fuzzy Inference* systems. The fundamental concept is to combine *fuzzy rule* logic with multi-layer mapping characteristic of neural networks.

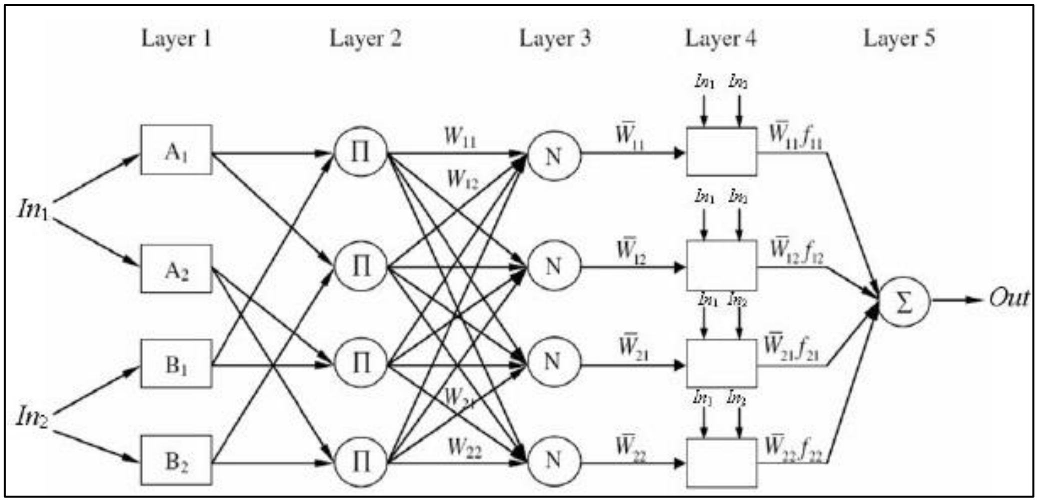


Fig. 2 An example of ANFIS architecture

In a naïve implementation, each input dimensions may have uniform partitions with a *membership function* situated at each partition. A partition is multi-dimensional space is thus interpretable as combination of per-dimension partitions. Certainly, if the network if fully connected, the combinations grow exponentially for high-dimensions. This makes learning intractable.

The ANFIS architecture for handling *one-hot encoded* data is trickier. The high dimensionality of the input-space makes it infeasible for fuzzy approximation using simple grid partitioning. Usually, Principle Component Analysis (PCA) or other dimensionality-reduction method might be useful here. However, the discreteness of categorical data makes it harder to perform such pre-processing. **PCA may or may not work for such data but it does not make sense to apply a *covariance* based method to categorical data.** However, **Subtractive Clustering** can implicitly exploit data similarity based on closeness and group the *fuzzy inference* partitions to make ANFIS training tractable. Moreover, since closeness estimation is done within hypercube bounding boxes, similarity should itself incorporate Manhattan distance which is suitable for such discrete data.

Subtractive clustering assumes that each data point is a potential cluster center. The algorithm does the following [3]:

1. Calculate the likelihood that each data point would define a cluster center, based on the density of surrounding data points.
2. Choose the data point with the highest potential to be the first cluster center.
3. Remove all data points near the first cluster center. The vicinity is determined using a hypercube bound.
4. Choose the remaining point with the highest potential as the next cluster center.
5. Repeat steps 3 and 4 until all the data is within the influence range of a cluster center.

This is an extension to Mountain Method [4].

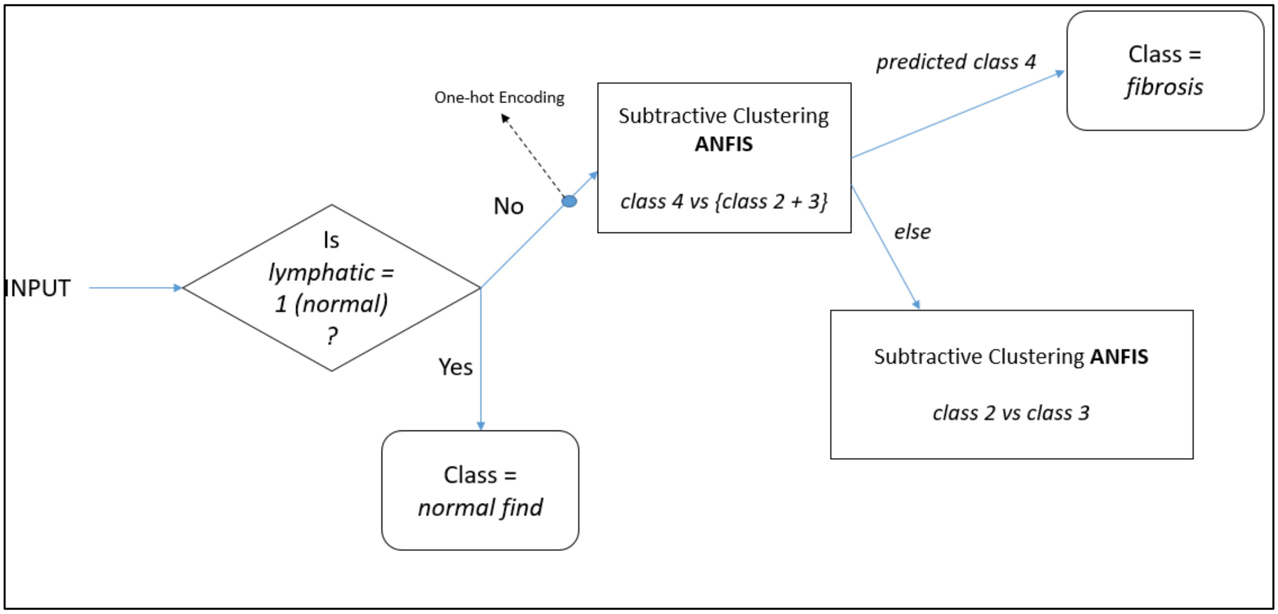


Fig. 3 Model – 2 Structure

IV. EXPERIMENT AND RESULTS

A. Data Split

Before training the dataset is partitioned into Training and Evaluation (Test) sets as follows (*approx. 80% - 20%),*

Class 1 *Normal find*: There are 2 data samples. Both are put into Evaluation set.

Class 2 *Metastases*: There are 81 data samples. 17 samples are put into Evaluation set.

Class 3 *Malign Lymph*: There are 61 data samples. 13 samples are put into Evaluation set.

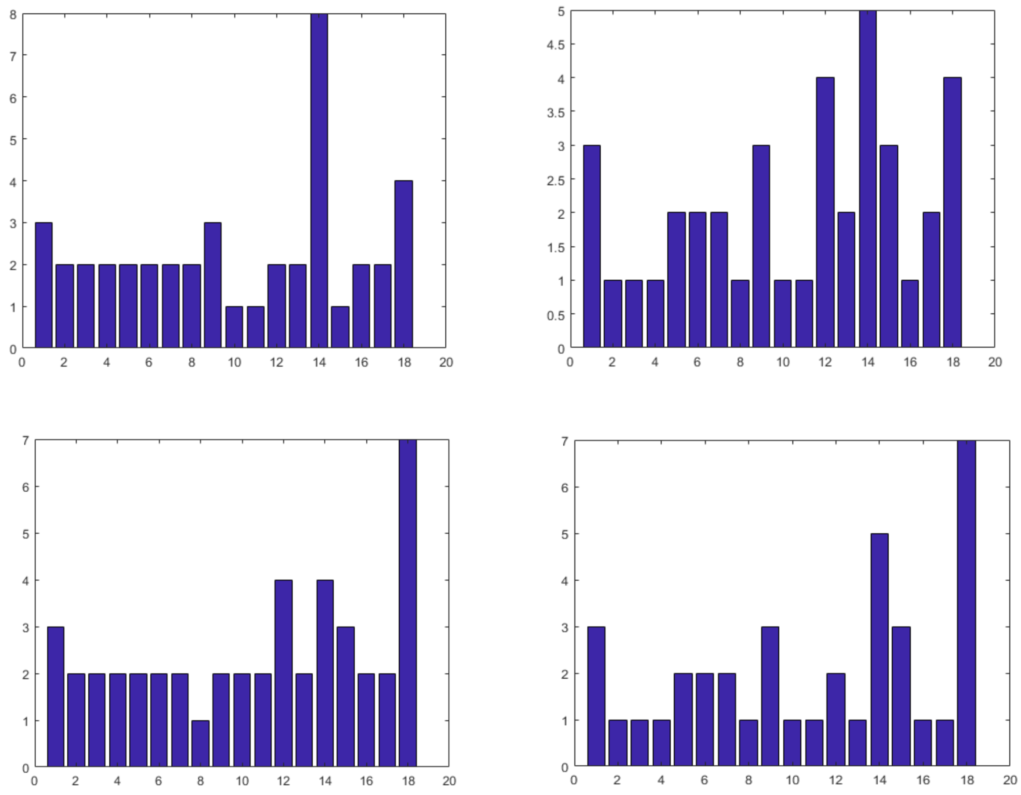
Class 4 *Fibrosis*: There are 4 data samples. 2 samples are put into Evaluation set.

TOTAL:

Training Samples = 114

Evaluation Samples = 34

The evaluation samples for Class 2 and Class 3 are chosen at random. The two evaluation samples for Class 4 are chosen by observing following distributions for the 18 attributes,



These plots show attribute values for the four data points (in clockwise sequence 1, 2, 3, 4) of *class 4*. Initial choice for training data is made based on variations – hence data sample 2 & 4 are chosen for training and 1 & 3 for evaluation. Further examination of the cosine distances (*multiplied by 10)* reveal:

Sample 1, *Sample* 2:	0.7777
Sample 1, *Sample* 3:	1.2623
Sample 1, *Sample* 4:	1.0481
Sample 2, *Sample* 3:	0.5595
Sample 2, *Sample* 4:	0.6141
Sample 3, *Sample* 4:	0.4802

The distances are scaled by a factor of 10. Apparently, sample 2 has low distances from other data points. Same applies for sample 4. Hence, these two were assessed to be able to provide relatively good generalization.

The training data is further divided into two sub-sets: (i) with all the samples of *class 4* removed; (ii) with all *class 2* and *class 3* samples merged into a single class. The former sub-set is for class 2 vs 3 classification, while the latter is for class 4 vs {2 + 3}.

B. Algorithm Settings

The RUSBoost Tree algorithm in Model-1 uses *Decision Trees* as basic learners. Maximum split is set as 20. The stage-3 module of Model-1 can use one of the three algorithms: *cosine kNN, Gaussian-kernel SVM, and Cubic-kernel SVM.* The performances are compared later. The cosine kNN uses k = 10 and equally weighted cosine distances. Cubic SVM is based on an order-3 polynomial kernel, and the Gaussian SVM based on Gaussian kernel with scale of 6.9. In Model-2, the ANFIS uses Subtractive Clustering with *Gaussian Membership Functions* to generate a fuzzy inference partitions. The training epochs value is specified as 20.

Post-training, the parameters of learned ANFIS structures are as follows:

ANFIS : class 4 vs class {2+3}	**ANFIS** : class 2 vs class 3
Number of nodes: 4850 Number of linear parameters: 2400 Number of nonlinear parameters: 4700 Total number of parameters: 7100 Number of fuzzy rules: 50	Number of nodes: 5618 Number of linear parameters: 2784 Number of nonlinear parameters: 5452 Total number of parameters: 8236 Number of fuzzy rules: 58

The size of the structure comes out to be quite large despite clustering. However, subtractive clustering is beneficial because ANFIS can at least be applied to the *one-hot encoded* features tractably.

C. Evaluation Results

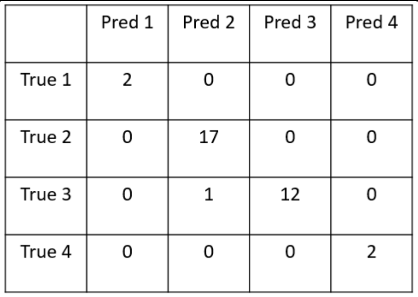
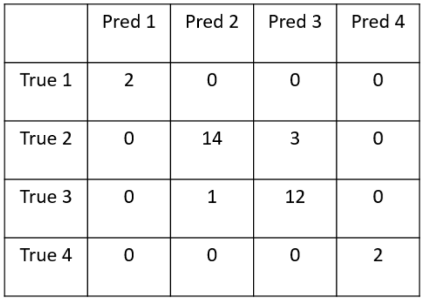
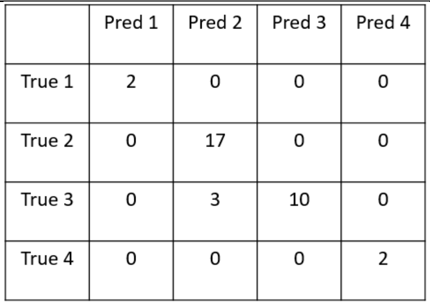
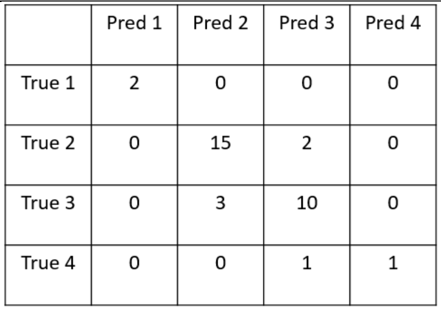
<table>
<tr><th rowspan="2">Architecture</th><th rowspan="2">Confusion Matrix
Total:
class 1 = **2**; class 2 = **17**
class 3 = **13;** class 4 = **2**</th><th colspan="4">**Per-class Error**</th></tr>
<tr><th>Class 1
out of 2</th><th>Class 2
out of 20</th><th>Class 3
out of 10</th><th>Class 4
out of 2</th></tr>
<tr><td>Model-1 with
RUSBoot and Cosine kNN</td><td>
<table>
<tr><th></th><th>Pred 1</th><th>Pred 2</th><th>Pred 3</th><th>Pred 4</th></tr>
<tr><td>True 1</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>True 2</td><td>0</td><td>17</td><td>0</td><td>0</td></tr>
<tr><td>True 3</td><td>0</td><td>1</td><td>12</td><td>0</td></tr>
<tr><td>True 4</td><td>0</td><td>0</td><td>0</td><td>2</td></tr>
</table></td><td>0</td><td>0</td><td>1</td><td>0</td></tr>
<tr><td>Model-1 with
RUSBoot and Gaussian SVM</td><td>
<table>
<tr><th></th><th>Pred 1</th><th>Pred 2</th><th>Pred 3</th><th>Pred 4</th></tr>
<tr><td>True 1</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>True 2</td><td>0</td><td>14</td><td>3</td><td>0</td></tr>
<tr><td>True 3</td><td>0</td><td>1</td><td>12</td><td>0</td></tr>
<tr><td>True 4</td><td>0</td><td>0</td><td>0</td><td>2</td></tr>
</table></td><td>0</td><td>3</td><td>1</td><td>0</td></tr>
<tr><td>Model-1 with
RUSBoot and Cubic SVM</td><td>
<table>
<tr><th></th><th>Pred 1</th><th>Pred 2</th><th>Pred 3</th><th>Pred 4</th></tr>
<tr><td>True 1</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>True 2</td><td>0</td><td>17</td><td>0</td><td>0</td></tr>
<tr><td>True 3</td><td>0</td><td>3</td><td>10</td><td>0</td></tr>
<tr><td>True 4</td><td>0</td><td>0</td><td>0</td><td>2</td></tr>
</table></td><td>0</td><td>0</td><td>3</td><td>0</td></tr>
<tr><td>Model-2 with **ANFIS**</td><td>
<table>
<tr><th></th><th>Pred 1</th><th>Pred 2</th><th>Pred 3</th><th>Pred 4</th></tr>
<tr><td>True 1</td><td>2</td><td>0</td><td>0</td><td>0</td></tr>
<tr><td>True 2</td><td>0</td><td>15</td><td>2</td><td>0</td></tr>
<tr><td>True 3</td><td>0</td><td>3</td><td>10</td><td>0</td></tr>
<tr><td>True 4</td><td>0</td><td>0</td><td>1</td><td>1</td></tr>
</table></td><td>0</td><td>2</td><td>3</td><td>1
50% error</td></tr>
</table>

Table 2. Performance of various models

V. DISCUSSION

Two hybrid architectures are discussed in this report for the purpose of classifying an imbalanced and discrete dataset of *nominal* and *ordinal* variables from Lymphography domain. Multi-stage hybrid structure is designed to handle the imbalanced classes in specific manner. The first model comprises of RUSBoost ensemble classifier to identify *fibrosis* class, whereas Cosine kNN, Gaussian SVM, and Cubic SVM are compared for two-class classification between *metastases* and *malign lymph* classes. Each of these algorithms are suitable for non-linear approximation in discrete data space generated by *one-hot* encoding of the categorical features. However, Cosine kNN seems more accurate as per the error profile provided in Table 2. The second model is a hybrid structure consisting of ANFIS neural architectures. Both models also comprise of an IF-ELSE filter to identify the *normal find* cases. The ANFIS architecture is applied for discrete categorical data by choosing the partition method as Subtractive Clustering. While results are generally good, fuzzy inference of ANFIS does not seem sufficient to achieve best result for the *fibrosis* class which may perhaps be treated as outlier.

REFERENCES

[1] Seiffert, Chris, et al. "RUSBoost: A hybrid approach to alleviating class imbalance." *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40.1 (2010): 185-197.

[2] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system." *IEEE transactions on systems, man, and cybernetics*23.3 (1993): 665-685.

[3] https://www.mathworks.com/help/fuzzy/subclust.html#bvm9zpz-5

[4] Yager, R. and D. Filev, "Generation of Fuzzy Rules by Mountain Clustering," *Journal of Intelligent & Fuzzy Systems*, Vol. 2, No. 3, pp. 209-219, 1994.