E Ö T V Ö S L O R Á N D T U D O M Á N Y E G Y E T E M
I N F O R M A T I K A I K A R
MESTERSÉGES INTELLIGENCIA TANSZÉK

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

Antika Das

Dátum/date: 2021.01.01.-2021.06.30.

KUTATÁSI JELENTÉS/RESEARCH REPORT

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
INFORMATIKAI KAR
MESTERSÉGES INTELLIGENCIA TANSZÉK

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

Task Description/Feladat leírása

Create a new method by using NIPGBoard to minimize the false negative rates, especially in dermatology dataset, which are too small for the proper learning of deep neural network.

Az NIPGBoard szoftver használatával fejlesz-szen ki egy módszert a fals negatív esetek csökkentésesére mélyhálós elemzések során, speciálisan rákos bőrsejeteket tartalmazó adatbázisra.

One of the research barriers within Bosch cooperation is when Bosch misses the informatic knowledge in engineering tasks and vice versa, and there is a need at Bosch to recognize the newest IT solutions, but this is impossible in many engineering fields. The solution to bridge this gap can be the development and testing of NIPGBoard. One such database is the classification of dermatology diseases, in deeper melanoma recognition. The task relevance from Bosch side is the following:
NIPGBoard integrate different algorithms and using it ensemble will providing new opportunities. Such a kind of task is to minimize the false negative cases whiten the classification. The task will be interesting if the failure detection is critical and the database is not enough big for the learning of the deep neural network.

- Use a deep neural network for classification.
- By the help of NipgBoard project the results of deep neural network classification into the 3D sphere.
- After the training with the help of the labels analyze the result of classification.
- Use different graph cutting to automatize the cluster searching method.
- Find those false negative clusters where quality is sufficient.
- Take out this samples from test and training dataset.
- Make correction on the remining poor clusters by using KIRA.
- Use step 3-6 again.
- Start the training again on the remaining dataset.

A Bosch együttműködés akadálya a Bosch számára szükséges különböző mérnöki szakmai tudás hiánya az informatikában és fordítva, a Boschnál az informatika legújabb vívmányainak ismerete kívánatos, de számos mérnöki területen nem lehetséges. A megoldás a NIPGBoard fejlesztése és tesztelése különböző adatbázisokon. Az egyik ilyen adatbázis a bőrbetegségek osztályozása és azon belül a melanóma felismerése. A feladat relevanciája a Bosch szempontjából a következő:
A NIPGBoard különböző algoritmusokat integrál, amelyek együttes felhasználása számos eddig fel nem tárt lehetőséget rejt. Ilyen feladat az osztályozás eredményén a fals negatív esetek lecsökkentése. A feladat érdekessé válik akkor, amikor a hibadetektálás kritikus, de az adatbázis nem elegendő a mély háló tökéletes betanítására.
Feladatok:

- Alkalmazza a mély hálós osztályozási módszert
- Vetítse le a NIPGBoard segítségével a mély háló mély reprezentációjának eredményeit 3D térbe
- A címkék segítségével vizsgálja meg a klaszterezés eredményeit

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

- Alkalmazzon gráfvágási módszereket a klaszterek keresésének automatizálására.
- Keressen olyan fals negatív klasztereket, amelyek minősége megfelelő.
- Az azokhoz tartozó tanító és tesztelő adatokat vegye ki az adathalmazból
- Javítsa a kevésbé jó klasztereket a Kira algoritmus segítségével
- Alkalmazza a 3.-6. lépéseket
- A megmaradó adatokon indítsa el a tanítást a legelejéről.

Vállalás/output 2021. június 30.:
2021.06.30:
Report about the results of decreasing false neg rate, maybe publication.
Tanulmány a fals negatív eredmények csökkentéséről, esetleg publikáció

Results/Eredmények

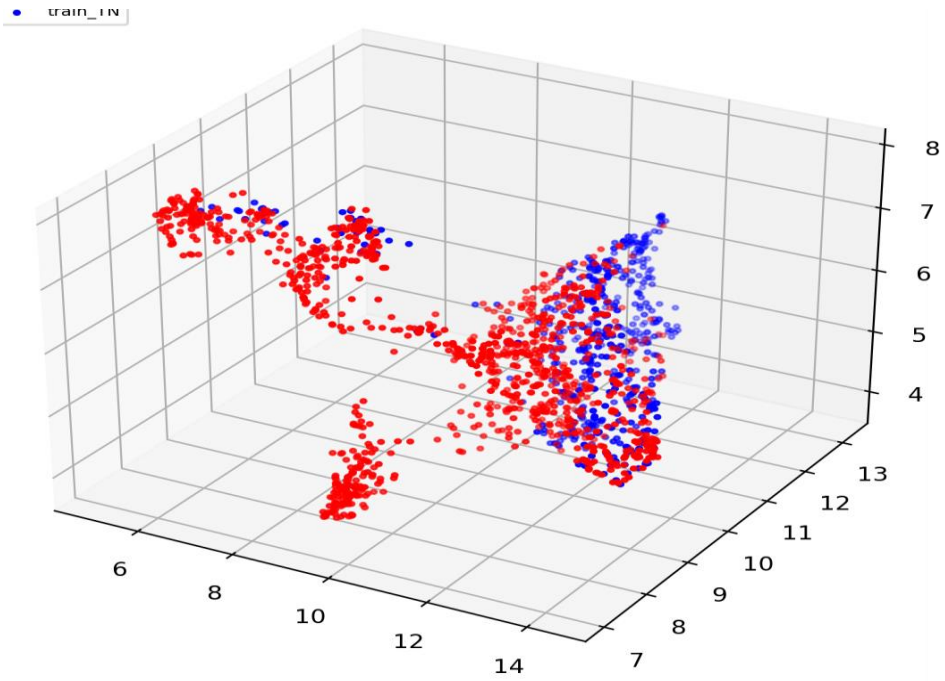
Outcomes, results, illustrations, tables, if relevant montly overview of your results

Results

Umap Visualization of subset train & subset_test :

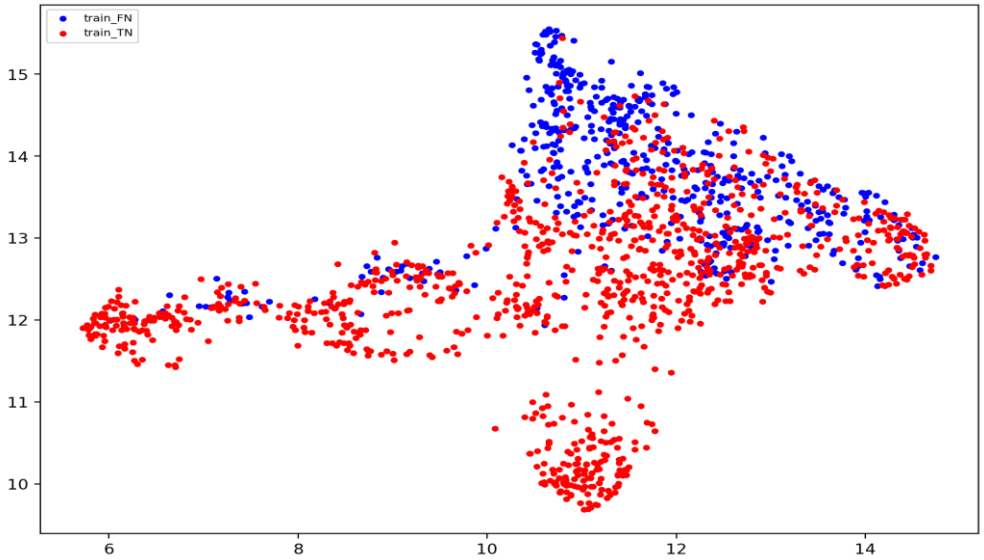
Note: This graph might slightly differ in each run of UMAP because of "random state".
Random state is a parameter for randomness both to speed up approximation steps, and to aid in solving hard optimization problems. But this randomness will not affect much into finding clusters using graph algorithms.

3D

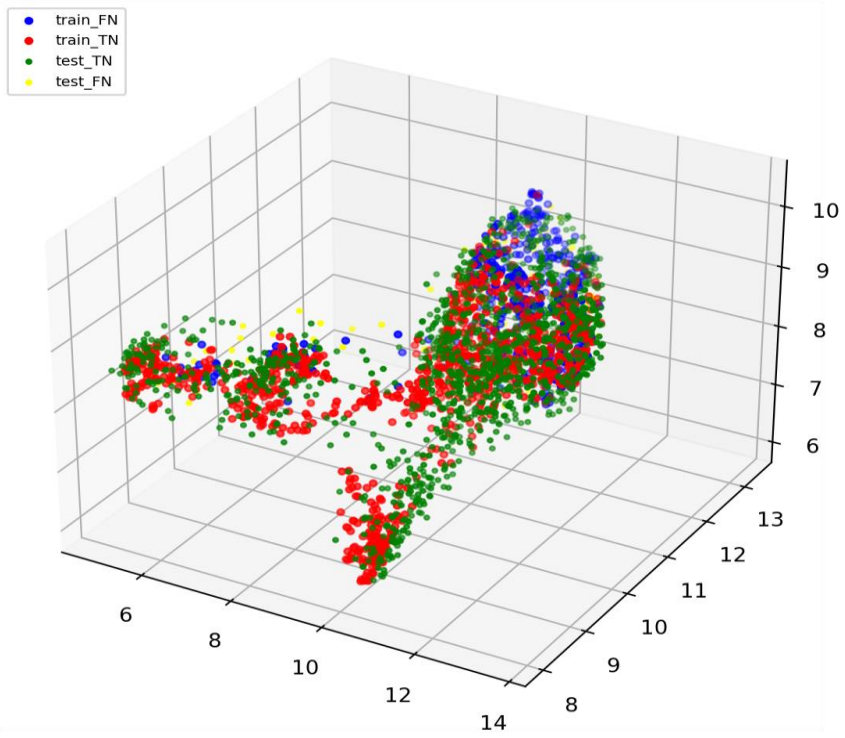


2D

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770



** Verify if test_embedding falls to the **same learned** space as train_embedding in 3D (unknown random state)

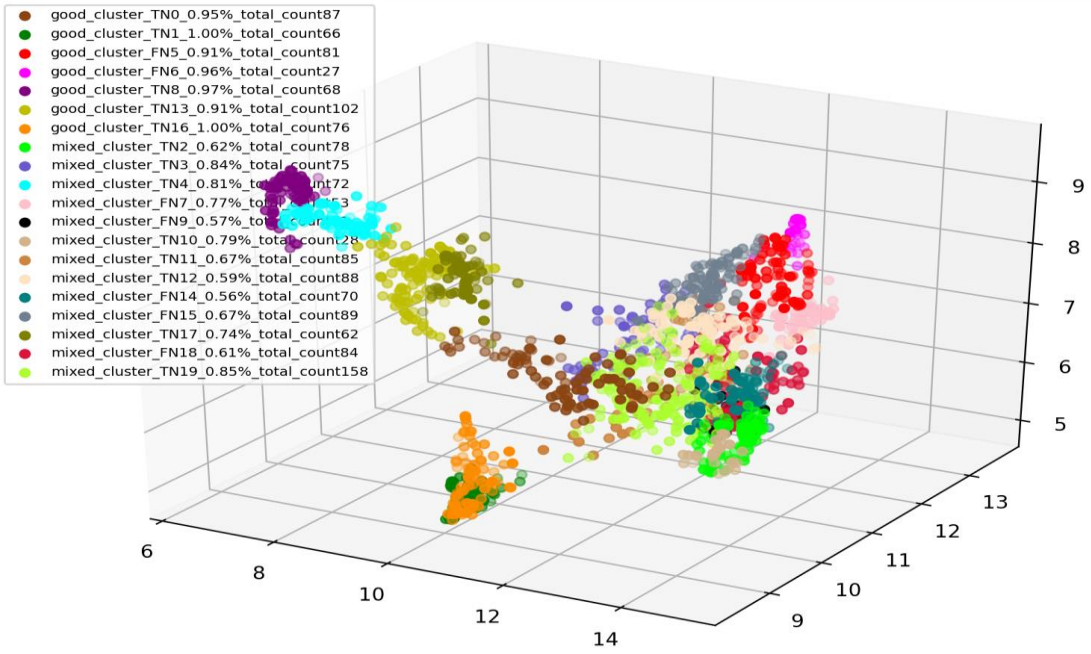


Louvain Community cluster analysis (Graphs)

*use the above umap as a input to the louvain algorithm to find out small clusters : total 20 small cluster detected

E Ö T V Ö S L O R Á N D T U D O M Á N Y E G Y E T E M
I N F O R M A T I K A I K A R
MESTERSÉGES INTELLIGENCIA TANSZÉK

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770



Cluster A, B, C

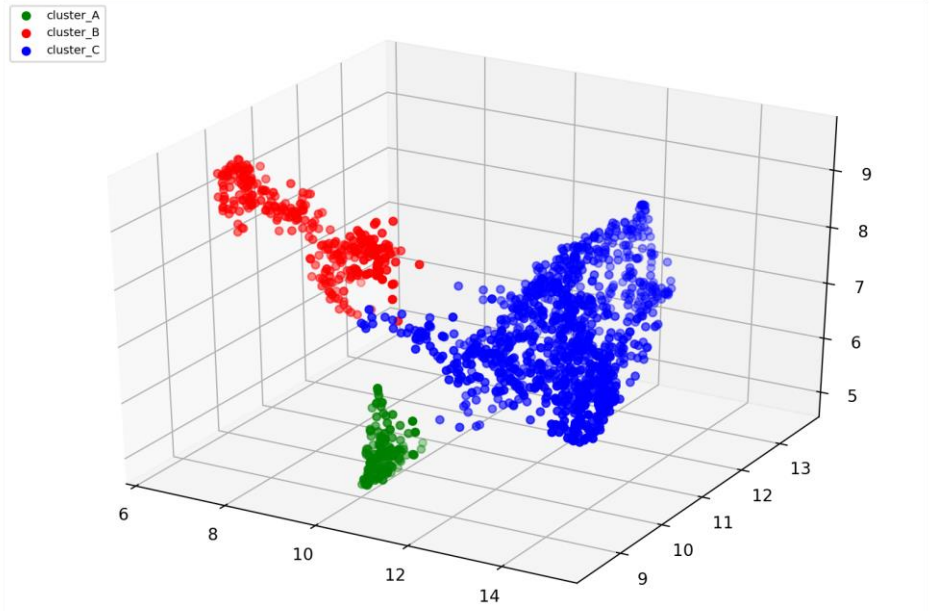
**compare the umap and Louvain output to decide which small clusters are to be merged together to get the final 3 cluster as shown in the below figure

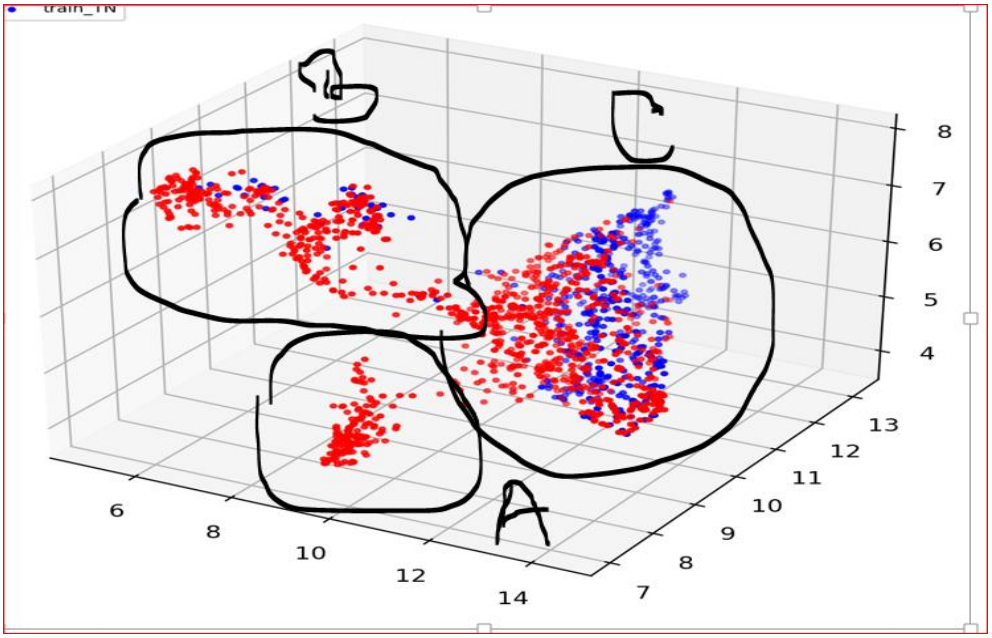
Cluster A =
"good_cluster_TN16_1.00%_total_count76"+"good_cluster_TN1_1.00%_total_count66"
Cluster B=
"good_cluster_TN8_0.97%_total_count68"+"mixed_cluster_TN4_0.81%_total_count72"+"good_cluster_TN13_0.91%_total_count102"+"mixed_cluster_TN17_0.74%_total_count62"

Cluster C= all remained small clusters

EÖTVÖS LORÁND TUDOMÁNYEGYETEM
INFORMATIKAI KAR
MESTERSÉGES INTELLIGENCIA TANSZÉK

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770





Test Samples Classification:

** Classify all the test embeddings to one of the three cluster using KNN techinque (details can be found on April,2021 monthy accomplished section)

A(TN_cluster): TN test: 105 FN test: 0

B(TN_cluster): TN test: 221 FN test: 24

C(FN_cluster + TN_cluster): TN test : 863 FN test: 85

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

Cluster A:

As cluster A is purely TN , also test samples classification proves the same we do not process the cluster further.

SVM (details can be found on May,2021 monthy accomplished section)

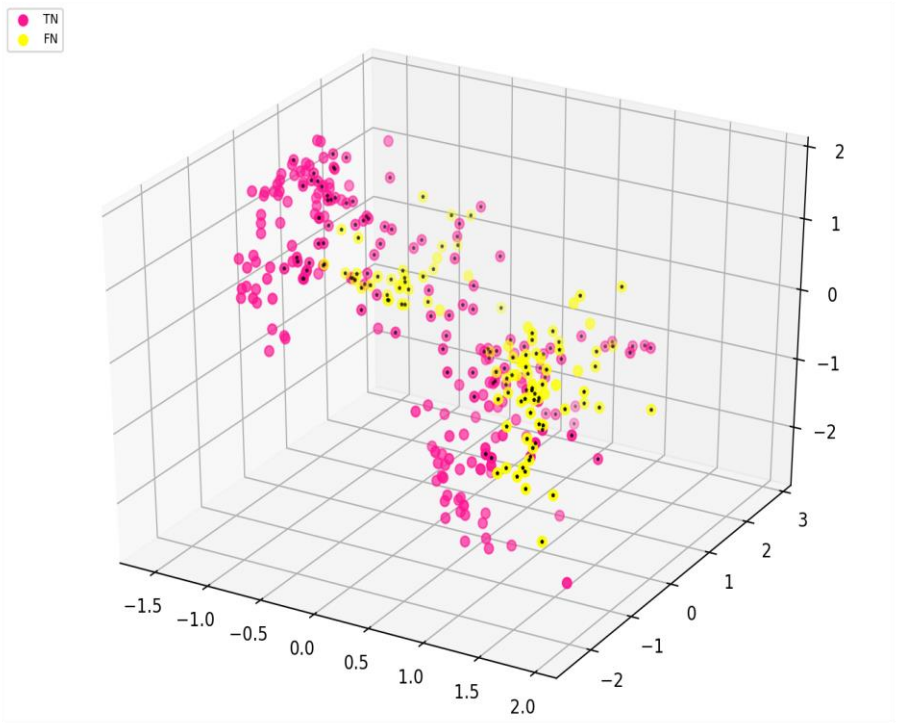
Output of cluster B SVM:

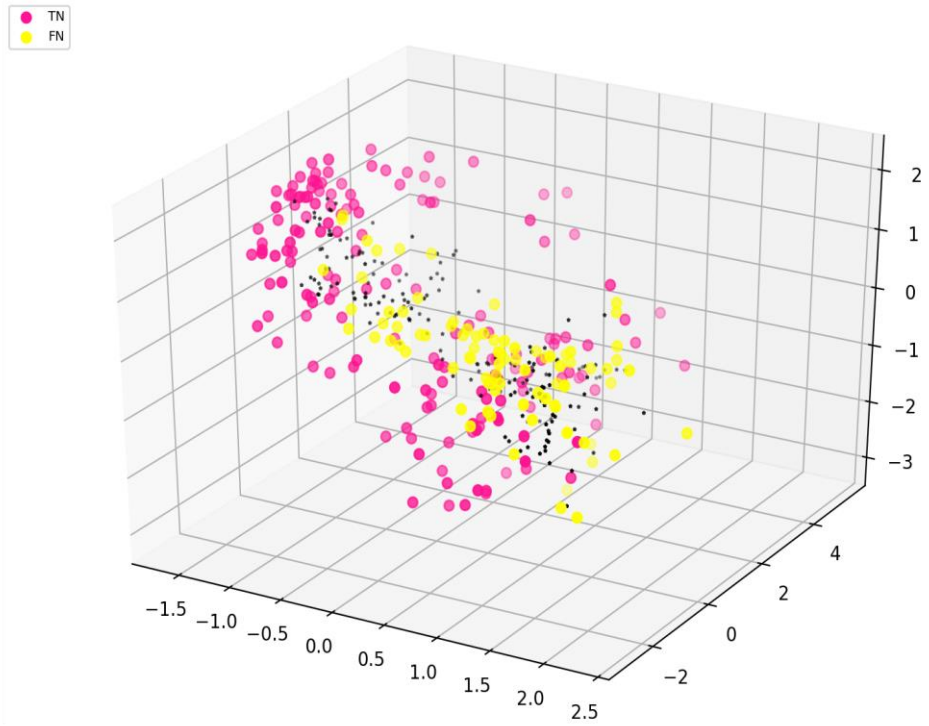
Unbalced data
C=10, gamma=0.1
weights ={0:6.5,1:1}
Train acc :0.7845683019567838
Test acc: 0.6541289592760181

Train Samples
Test samples

E Ö T V Ö S L O R Á N D T U D O M Á N Y E G Y E T E M
I N F O R M A T I K A I K A R
MESTERSÉGES INTELLIGENCIA TANSZÉK

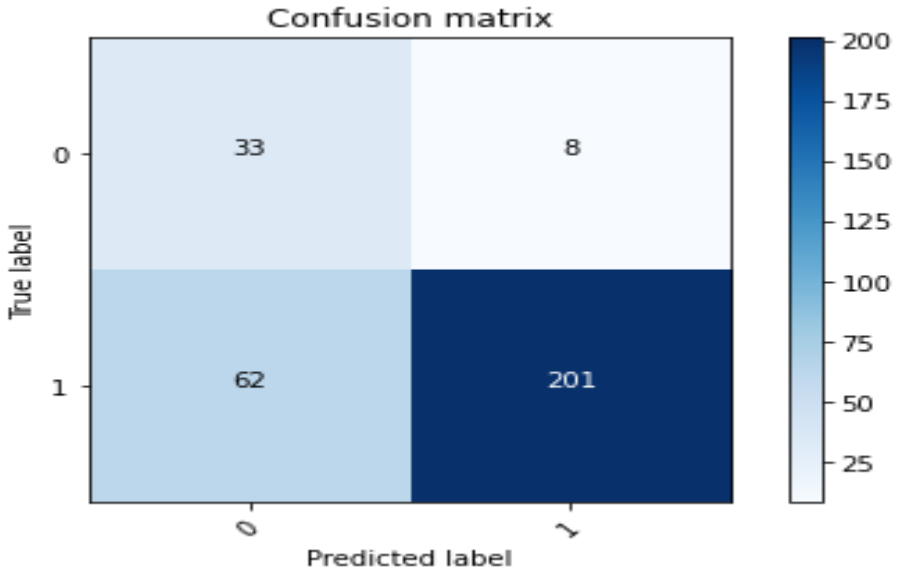
1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

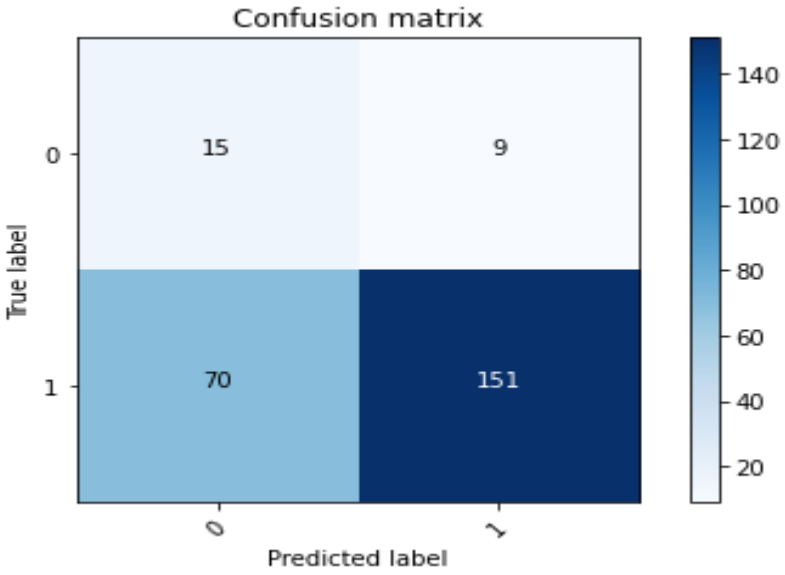




E Ö T V Ö S L O R Á N D T U D O M Á N Y E G Y E T E M
I N F O R M A T I K A I K A R
MESTERSÉGES INTELLIGENCIA TANSZÉK

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770





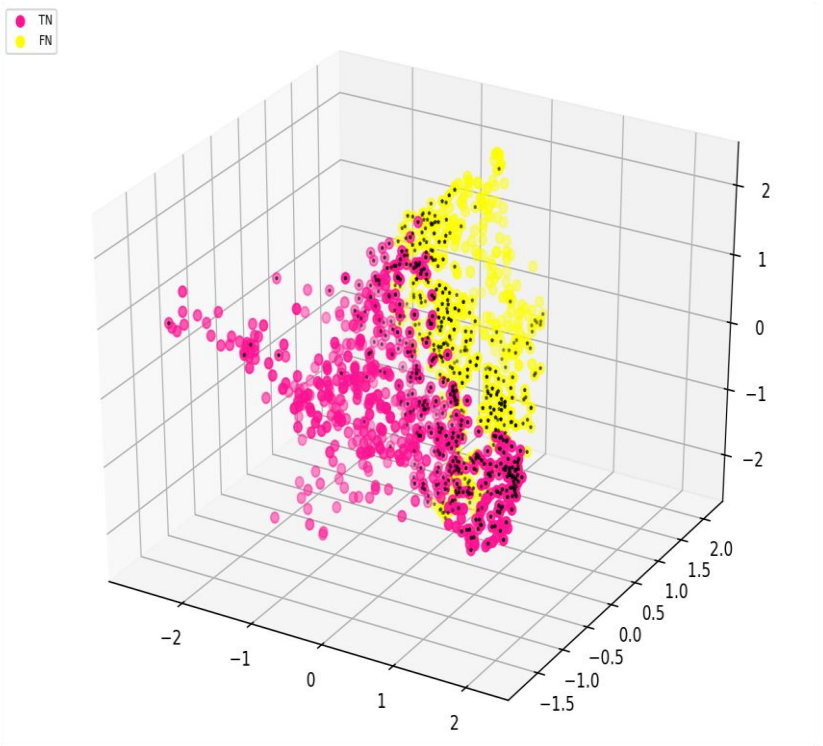
Output of cluster C_SVM:

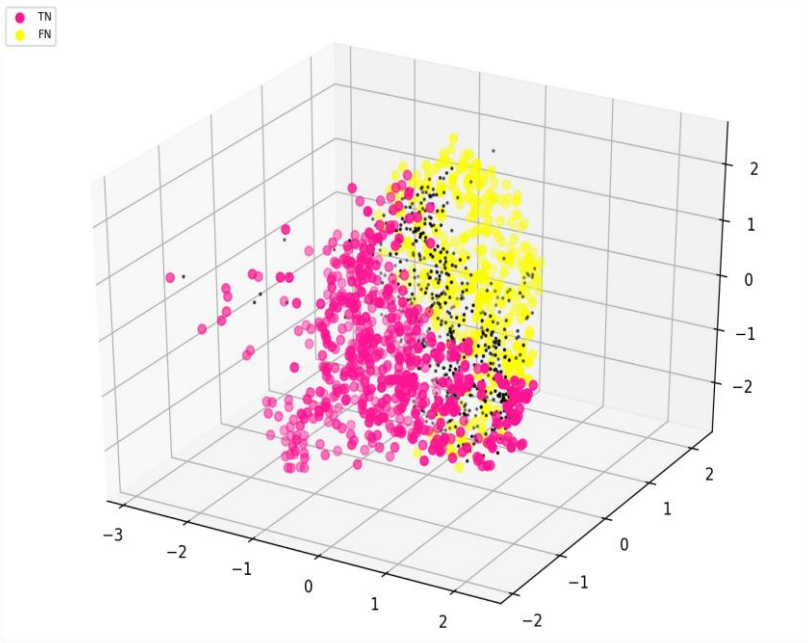
C=10, gamma=0.1
weights ={0:1,1:1}
Train acc: 0.775412387177093
Train acc: 0.713441483198146

Train Samples
Test samples

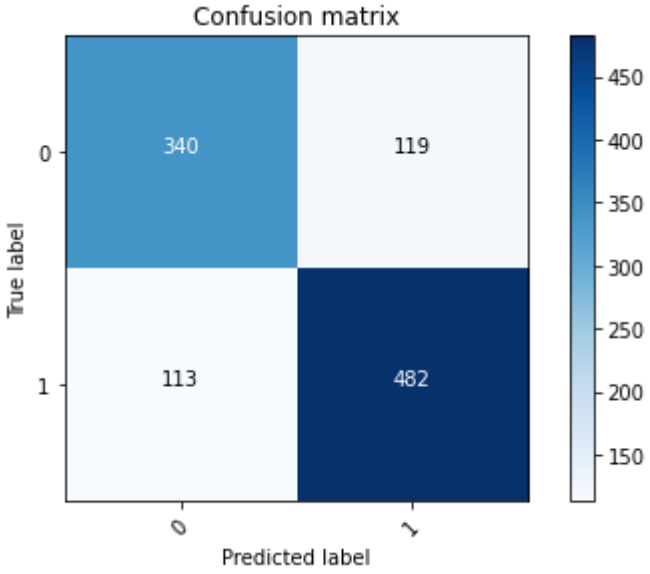
E Ö T V Ö S L O R Á N D T U D O M Á N Y E G Y E T E M
I N F O R M A T I K A I K A R
MESTERSÉGES INTELLIGENCIA TANSZÉK

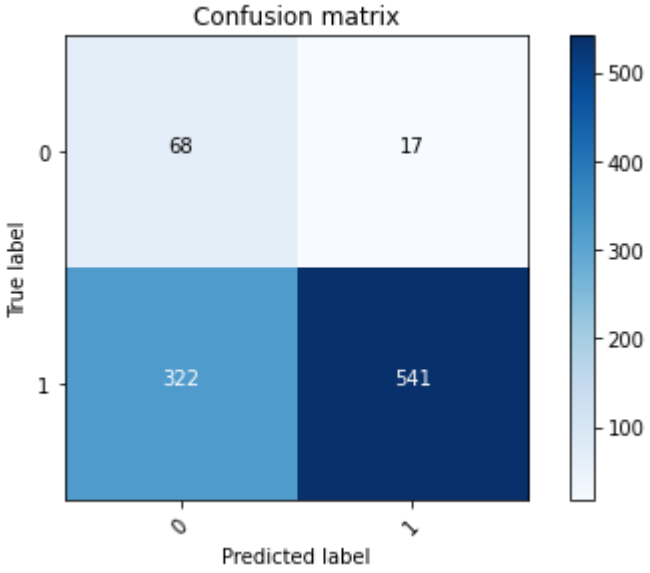
1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770





1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770





Executive summary/Összegzés

Overview of whole work with accomplished deadlines, monthly schedule

Accomplished tasks and monthly separation:

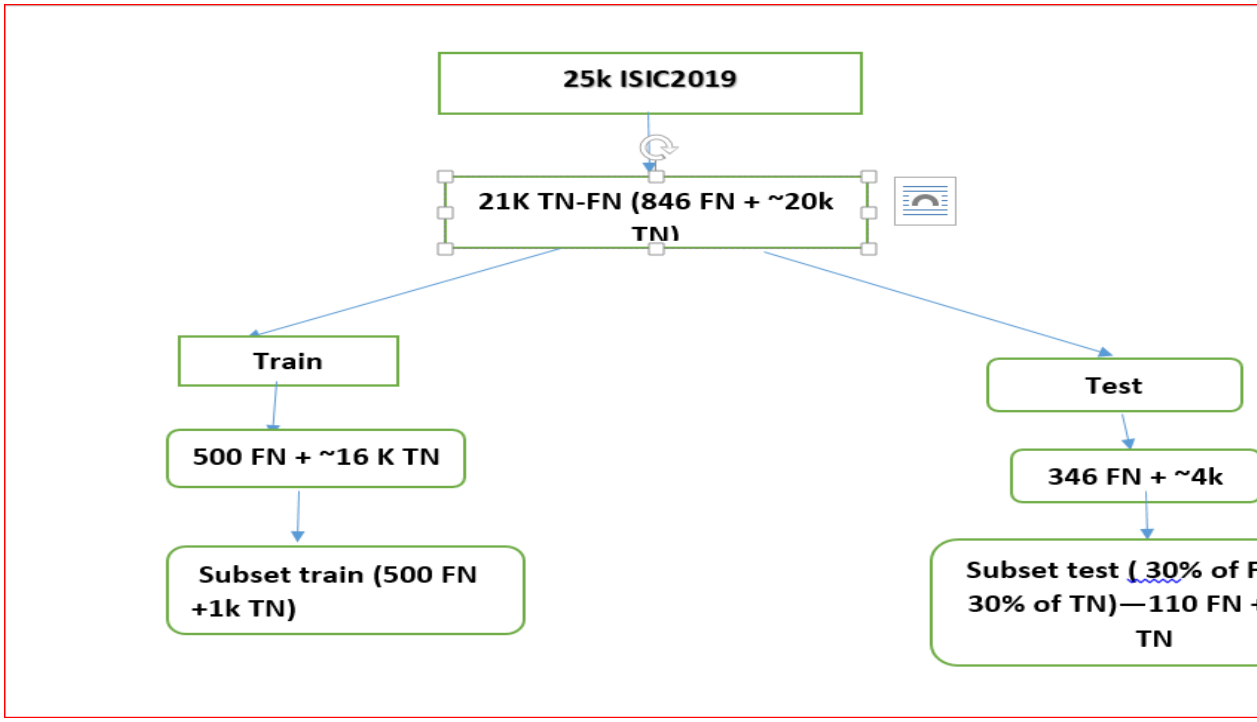
January, 2021

Data Preaparation & Feature Extraction (Embedding)

- Received necessary informations, pytorch models, weight files, ground truth,predictions etc. according to the plan.

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

- Efficient-net pytorch model was used to get the embedding or feature extraction for further processing.
- Data Preparation: Decide on threshold value using ROC curve analysis (55%). Convert the probability predictions to binary number using the threshold 55%. next comapre the ground truth and predictions to filter out the True Negative(TN) and False Negative(FN). Prediction==0 and ground truth==0 means TN, Prediction==0 and ground truth==1 means FN.
- Spilt tall the FN-TN into train and test and take subset.



- Input the subset train and subset test images to pre-trained pytorch model to get the embedding.
- This embeddings are used for dimension reduction algorithm such as PCA, UMAP.
- In this experiment the we have focused on 3D UMAP.
- So the 1792D vector embeddings will be reduced to 3D.
- UMAP parameters plays a vital role in achieving the final 3D vector, such as neighbouhood,min-distance, epochs, transform seed etc.

February, 2021

Docker,Environment Set up and Umap

- Set up Docker, solve docker related issues.
- Set up jupyter notebook and nipgboard in docker.

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

- Decide on Umap parameters, reduce the originanal embedding (1792) to 3D. Parameters : neighbouhood-43, epoch -500, min_dist=0.1,distance=”eucledian”.
- Once the we get the 3D embedding for train dataset use the same embedding space to transform the test embedding.

March, 2021

Umap and graph algorithm

- Once we had train and test 3D umap projection points we used them for further processing into graph algorithm – Louvain community detection algorithm.
- We input the train umap projections to the Louvain algorithm to get small clusters, the the resolution parameter is most important as it is responsibe for changing the size of the communities, it decides the boundaries of the community, increasing the resolution will increase the boundary or size of the community thus decreasing the cluster number and vice versa. Thus the number of clusters will decrease and increase accordingly.
- Once the cluster informations are achieved , we categorized the clusters into good and mixed clusters.Louvian algorithm gives two types of accuracy information, accuracy/percentage of each clusters detected by it and average estimated accuracy of all clusters. Now we decide a threshod value, if the accuarcy of each cluster is > threshold then we will consider that cluster as good cluster or else mixed cluster.
- One of the most important step after this is to merge the small clusters into 3/4 big clusters. For this a lot of visual inspection is needed , such as check nipgboard, check umap and louvain cluster output. This is a point where human knowledge and graph outputs are combined.

April, 2021

Cluster A,B,C

- Decide the final small clusters to be mereged , then merge the clusters to get back three big clusters namely A,B,C.
- Once we have this three cluters now classify the test samples into cluster A,B,C.Used KNN (K=1,distance=euclidean) to find out for each test sample which is the closest point in the train sample. Then the test sample will receive the domain of the cluster where the closet train point belongs to. e.g- Let's say I have a test sample P and I know the ground truth of it (lets say FN) now I

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

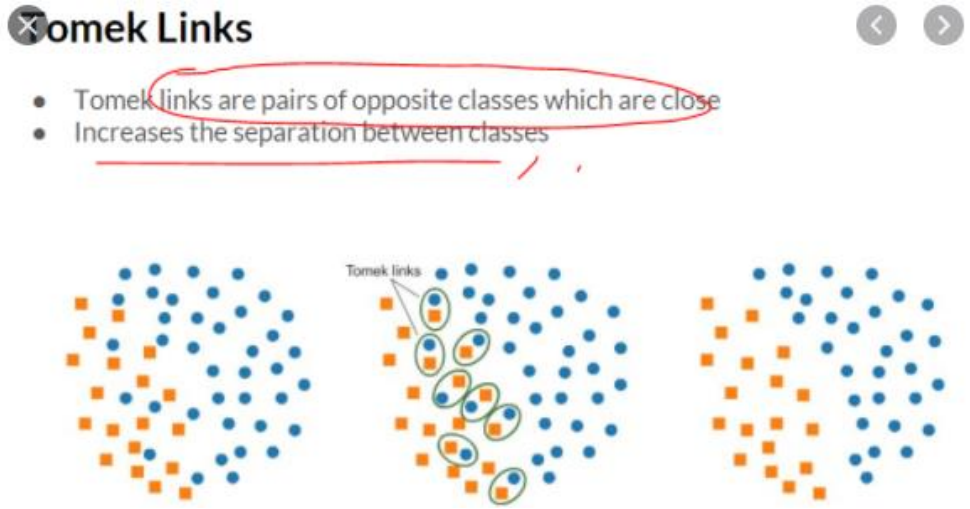
checked which train sample is closet to P, let's say Q (train sample), I already have the information that Q belongs to which cluster/domain lets say TN_cluster5 , that is how I can decide what should be the cluster/domain of the point P. So if Q belongs to a high-quality TN/FN or low quality TN/FN cluster then P belongs to the same cluster (in this case TN_cluster5) as well and receives the label as TN.

- Now treat the three cluster differently. Cluster A is a pure (100%) TN cluster, cluster B is mostly pure except a small portion of FN-the reason behind that, after investigation we found out that B cluster is effected by scale. C clusters containes most of the images from the subset train and its highly mixed - almost 50-50 TN-FN. So our main focus will be cluster C and seperate TN and FN.

May, 2021

SVM

- Use SVM to seperate the TN-FN points from cluster B,C.
- Most important parameters of SVM are kernel and slack variable. slack variable can be seen as a penalize parameter for miss-classification. Large values of C correspond to large error penalties while we are less strict about miss-classification errors if we choose smaller values for C. Large value of C means wide margin, small value of C means wide margin.
- Used Under-sampling: Tomek links, Tomek links are pairs of very close instances but of opposite classes. Removing the instances of the majority class of each pair increases the space between the two classes, facilitating the classification process.Tomek's link exists if the two samples are the nearest neighbors of each other.



- Final output of SVM (cluster B):

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770

Unbalced data
C=10, gamma=0.1
weights ={0:6.5,1:1}
Train acc :0.7845683019567838
Test acc: 0.6541289592760181

Balaced Data
SVM is overfitting (Even though with balanced the the train samples are seperated well with linear kernel, its not working well on test data)

- Final output of SVM (cluster C)

C=10, gamma=0.1
weights ={0:1,1:1}
Train acc: 0.775412387177093
Train acc: 0.713441483198146

June, 2021

Support vectors, ISIC_Combined, Documentaion

- Use the support vectors of cluster C, use the support vectors for further processing into kira. Use automatic pairing of kira on support vectors.—uncloncluded
- Did the same steps as above with isic_combined data, but the results were not satisfactory as the FN rate was already very low (0.07%).
- Making and organisingDocument.

E Ö T V Ö S L O R Á N D T U D O M Á N Y E G Y E T E M
I N F O R M A T I K A I K A R
MESTERSÉGES INTELLIGENCIA TANSZÉK

1117 Budapest, Pázmány Péter sétány 1/A., 7.em/50. Tel: 372-2500/6770