CSE431 - Natural Language Processing with Machine Learning

2nd Project

Salih Can AYDOĞDU-191805061

#Contains both the first and second parts of the project#

1. Introduction to the Project

The Kaggle link to get the dataset used in our project is as follows:
https://www.kaggle.com/tboyle10/medicaltranscriptions

The aim of this project is to accurately classify medical specialties based on transcript text. That is, to develop a system or model that, by analyzing the content of a particular piece of text, can accurately determine which medical specialty the text belongs to. This type of system can automatically classify health-related texts and direct them to the relevant medical specialty. If the project is successful, this type of technology could help manage medical documents and transcripts more effectively and provide rapid access to healthcare professionals.

2. Examining the data I have

As I see, this dataset contains medical transcriptions. When I examine the first 5 lines;

Unnamed: 0: This is a column containing line numbers, used for indexing purposes.

description: This is the column that contains the description of the medical event or situation. For example, "A 23-year-old white female presents with comp...".

medical_specialty: This is the column that indicates the medical specialty. For example, "Allergy / Immunology", "Bariatrics", "Cardiovascular / Pulmonary".

sample_name: This is the column that specifies the name of the sample. For example, "Allergic Rhinitis", "Laparoscopic Gastric Bypass Consult - 2".

transcription: This is the column containing the medical transcription text. For example, "SUBJECTIVE:, This 23-year-old white female pr...".

keywords: This is the column containing keywords. For example, "allergy / immunology, allergic rhinitis, aller...".

When trying to conduct an analysis using machine learning methods on this data set, some difficulties may be encountered. First of all, the text data in the "description" and "transcription" columns appear to contain a variety of medical terms, symbols, and formats. It is obvious that pre-processing steps are needed to create a certain order and standard in the data set. Therefore, it is obvious that preprocessing steps may need to be applied to understand and process the data. (It is worth noting that since I will benefit from the "transcription" feature rather than the "description" column in this project, preprocessing will be done through this feature.)

3. Reducing noise in the data and applying preprocessing operations

The preprocessing steps performed in this step helped the text data become more consistent and rich in meaning. . These steps reduce noise in the data set and make the analysis more reliable.

Additionally, it ensures that the dataset is suitable for classification algorithms to be applied to accurately classify medical specialties.

1. Obtaining medical specialties (categories) in the data set and removing those with feIr than 50 examples

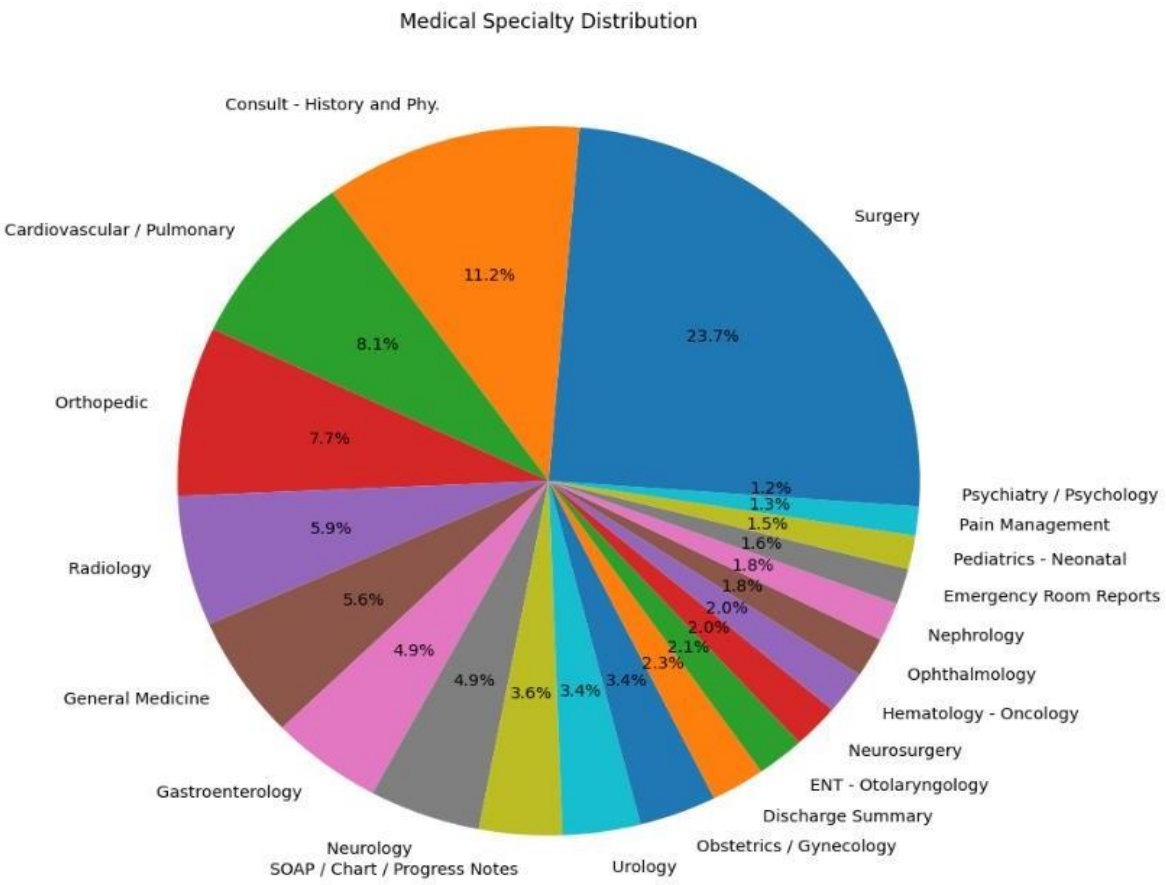
```
----------Original Categories ------------------------
Cat:1  Allergy / Immunology : 7
Cat:2  Autopsy : 8
Cat:3  Bariatrics : 18
Cat:4  Cardiovascular / Pulmonary : 371
Cat:5  Chiropractic : 14
Cat:6  Consult - History and Phy. : 516
Cat:7  Cosmetic / Plastic Surgery : 27
Cat:8  Dentistry : 27
Cat:9  Dermatology : 29
Cat:10  Diets and Nutritions : 10
Cat:11  Discharge Summary : 108
Cat:12  ENT - Otolaryngology : 96
Cat:13  Emergency Room Reports : 75
Cat:14  Endocrinology : 19
Cat:15  Gastroenterology : 224
Cat:16  General Medicine : 259
Cat:17  Hematology - Oncology : 90
Cat:18  Hospice - Palliative Care : 6
Cat:19  IME-QME-Work Comp etc. : 16
Cat:20  Lab Medicine - Pathology : 8
Cat:21  Letters : 23
Cat:22  Nephrology : 81
Cat:23  Neurology : 223
Cat:24  Neurosurgery : 94
Cat:25  Obstetrics / Gynecology : 155
Cat:26  Office Notes : 50
Cat:27  Ophthalmology : 83
Cat:28  Orthopedic : 355
Cat:29  Pain Management : 61
Cat:30  Pediatrics - Neonatal : 70
Cat:31  Physical Medicine - Rehab : 21
Cat:32  Podiatry : 47
Cat:33  Psychiatry / Psychology : 53
Cat:34  Radiology : 273
Cat:35  Rheumatology : 10
Cat:36  SOAP / Chart / Progress Notes : 166
Cat:37  Sleep Medicine : 20
Cat:38  Speech - Language : 9
Cat:39  Surgery : 1088
Cat:40  Urology : 156
----------------------------------
```

As you can see, there are 40 different areas of expertise in our dataset.

When I filter the specialties that have more than 50 examples in the data set (more than 50 labels)

```
============Reduced Categories =====================
Cat:1  Cardiovascular / Pulmonary : 371
Cat:2  Consult - History and Phy. : 516
Cat:3  Discharge Summary : 108
Cat:4  ENT - Otolaryngology : 96
Cat:5  Emergency Room Reports : 75
Cat:6  Gastroenterology : 224
Cat:7  General Medicine : 259
Cat:8  Hematology - Oncology : 90
Cat:9  Nephrology : 81
Cat:10  Neurology : 223
Cat:11  Neurosurgery : 94
Cat:12  Obstetrics / Gynecology : 155
Cat:13  Ophthalmology : 83
Cat:14  Orthopedic : 355
Cat:15  Pain Management : 61
Cat:16  Pediatrics - Neonatal : 70
Cat:17  Psychiatry / Psychology : 53
Cat:18  Radiology : 273
Cat:19  SOAP / Chart / Progress Notes : 166
Cat:20  Surgery : 1088
Cat:21  Urology : 156
============ Reduced Categories =====================
```

This number drops to 21.



Distribution of medical specialties with more than 50 labels

2. Removing rows with empty values in the "Transcription" feature

3. Cleaning and lemmatizing the "Transcription" feature

Let's look at an example of a row in the original dataset before cleaning and lemmatizing operations are performed.

Sample Transcription 1:CC:, Confusion and slurred speech.,HX , (primarily obtained from boyfriend): This 31 y/o RHF experienced a "flu-like illness 6-8 Ieks prior to presentation. 3-4 Ieks prior to presentation, she was found "passed out" in bed, and when awoken appeared confused, and lethargic. She apparently recovered within 24 hours. For two Ieks prior to presentation she demonstrated emotional lability, uncharacteristic of her (outbursts of anger and inappropriate laughter). She left a stove on.,She began slurring her speech 2 days prior to admission. On the day of presentation she developed right facial Iakness and began stumbling to the right. She denied any associated headache, nausea, vomiting, fever, chills, neck stiffness or visual change. There was no history of illicit drug/ETOH use or head trauma.,PMH:, Migraine Headache.,FHX: , Unremarkable.,SHX:
,Divorced. Lives with boyfriend. 3 children alive and Ill. Denied tobacco/illicit drug use. Rarely consumes ETOH.,ROS:, Irregular menses.,EXAM: ,BP118/66. HR83. RR 20. T36.8C.,MS: Alert and oriented to name only. Perseverative thought processes. Utilized only one or two word ansIrs/phrases. Non-fluent. Rarely folloId commands. Impaired writing of name.,CN: Flattened right nasolabial fold only.,Motor: Mild Iakness in RUE manifested by pronator drift. Other extremities Ire full strength.,Sensory: withdrew to noxious stimulation in all 4 extremities.,Coordination: difficult to assess.,Station: Right pronator drift.,Gait: unremarkable.,Reflexes: 2/2BUE, 3/3BLE, Plantars Ire flexor bilaterally.,General Exam: unremarkable.,INITIAL STUDIES:, CBC, GS, UA, PT, PTT, ESR, CRP, EKG Ire all unremarkable. Outside HCT shoId hypodensities in the right putamen, left caudate, and at several subcortical locations (not specified).,COURSE: ,MRI Brian Scan, 2/11/92 revealed an old lacunar infarct in the right basal ganglia, edema within the head of the left caudate nucleus suggesting an acute ischemic event, and arterial enhancement of the left MCA distribution suggesting slow flow. The latter suggested a vasculopathy such as Moya Moya, or fibromuscular dysplasia.
HIV, ANA, Anti-cardiolipin Antibody titer, Cardiac enzymes, TFTs, B12, and cholesterol studies Ire unremarkable.,She underInt a cerebral angiogram on 2/12/92. This revealed an occlusion of the left MCA just distal to its origin. The distal distribution of the left MCA filled on later films through collaterals from the left ACA. There was also an occlusion of the right MCA just distal to the temporal branch. Distal branches of the right MCA filled through collaterals from the right ACA. No other vascular abnormalities Ire noted. These findings Ire felt to be atypical but nevertheless suspicious of a large caliber vasculitis such as Moya Moya disease. She was subsequently given this diagnosis. Neuropsychologic testing revealed widespread cognitive dysfunction with particular impairment of language function. She had long latencies responding and understood only simple questions. Affect was blunted and there was distinct lack of concern regarding her condition. She was subsequently discharged home on no medications.,In 9/92 she was admitted for sudden onset right hemiparesis and mental status change. Exam revealed the hemiparesis and in addition she was found to have significant neck lymphadenopathy. OB/GYN exam including cervical biopsy, and abdominal/pelvic CT scanning revealed stage IV squamous cell cancer of the cervix. She died 9/24/92 of cervical cancer.

The operations that occur during the text cleaning process are as follows;

#remove all punctuation marks from the text
#remove all digits from text
#convert text to loIrcase letters
#changing special characters to spaces (such as semicolons, square brackets)

Another process, lemmatization, means going to the root of the word.

After applying the clearing and lemmatize processes, the sample text turned into this;

Sample Transcription 1:cc confusion and slurred speech hx primarily obtained from boyfriend this yo rhf experienced a flulike illness Iek prior to presentation obgyn exam including cervical biopsy and abdominalpelvic ct scanning revealed stage iv squamous cell cancer of the cervix

4. Performing Bag-of-Words (CountVectorizer) and TF-IDF (TfidfVectorizer) in the "Transcription" feature

Bag-of-Words (BoW) is a text mining and natural language processing technique that represents text by counting word frequencies. CountVectorizer, a BoW method, converts text into a numerical vector, treating each word as a feature with its frequency.

TF-IDF, short for "Term Frequency-Inverse Document Frequency," is widely used in text mining and language processing. It assesses a word's importance in a document, finding applications in text mining, information extraction, and text classification.

Visualized version of the matrix obtained from Bag-of-Words (CountVectorizer) and TF-IDF (TfidfVectorizer) methods by reducing it to two dimensions

5. Applying PCA (Principal Component Analysis) to the resulting matrix

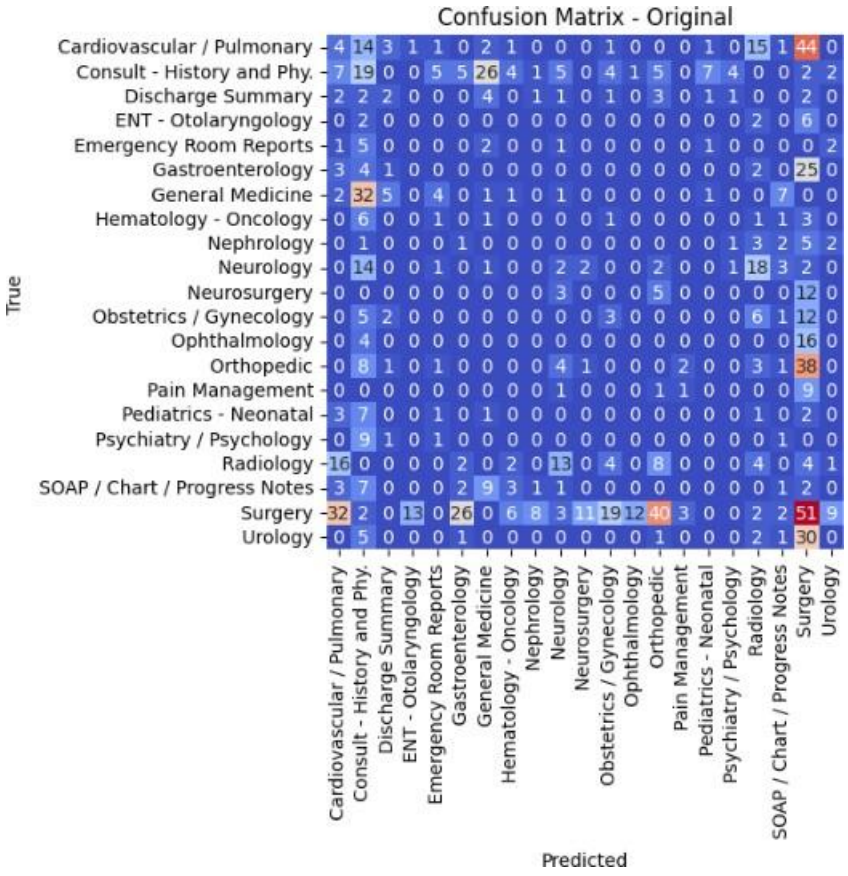
PCA performs dimensionality reduction, especially when used in high-dimensional data sets. This reduces the number of features contained in the data, thus allowing the model to perform better, reducing the risk of overfitting and reducing computational cost. The main purpose of PCA is to make the data more effective by preserving the variability in the data but expressing it with feIr features.

We continued our operations by applying PCA to the matrix I obtained in the previous stage.

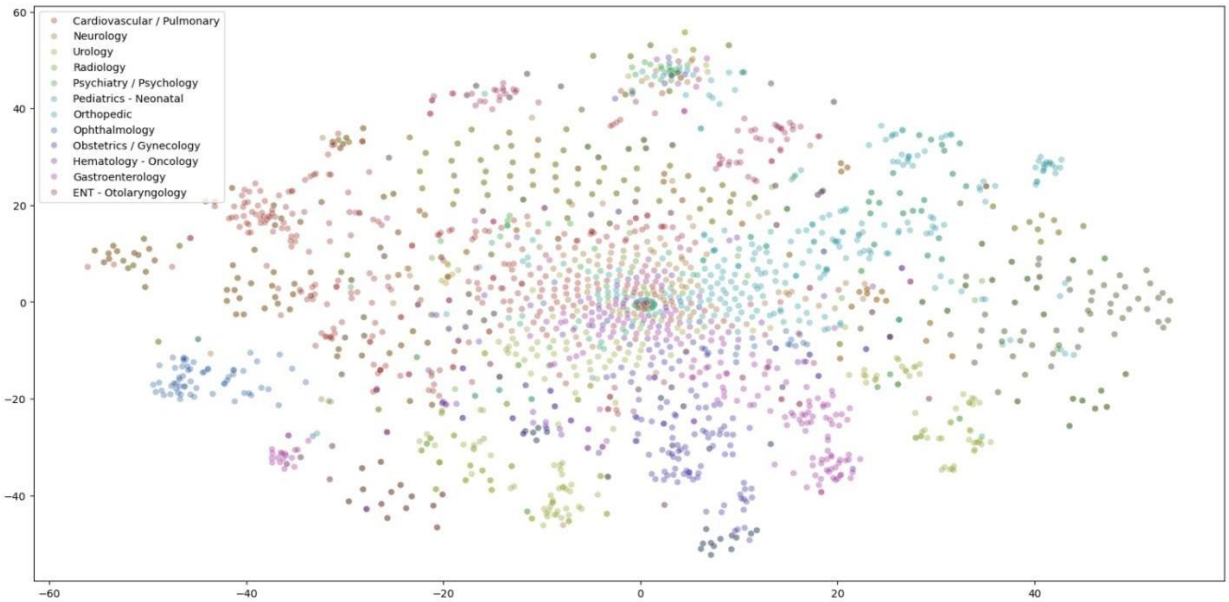
After applying these 5 steps, I wanted to check our machine learning metrics using the Random Forest classification method().

Category	Sampler	Accuracy	Precision	Recall	F1 Score
Cardiovascular / Pulmonary	Original	0.045454545454545456	0.045454545454545456	0.045454545454545456	0.045454545454545456
Cardiovascular / Pulmonary	SMOTE	0.4861111111111111	0.4861111111111111	0.4861111111111111	0.4861111111111111
Consult - History and Phy.	Original	0.1958762886597938	0.1958762886597938	0.1958762886597938	0.1958762886597938
Consult - History and Phy.	SMOTE	0.2054794520547945	0.2054794520547945	0.2054794520547945	0.2054794520547945
Discharge Summary	Original	0.1	0.1	0.1	0.10000000000000002
Discharge Summary	SMOTE	0.8312236286919831	0.8312236286919831	0.8312236286919831	0.8312236286919831
ENT - Otolaryngology	Original	0.0	0.0	0.0	0.0
ENT - Otolaryngology	SMOTE	0.8227272727272728	0.8227272727272728	0.8227272727272728	0.8227272727272728
Emergency Room Reports	Original	0.0	0.0	0.0	0.0
Emergency Room Reports	SMOTE	0.9004524886877828	0.9004524886877828	0.9004524886877828	0.9004524886877829
Gastroenterology	Original	0.0	0.0	0.0	0.0
Gastroenterology	SMOTE	0.6652173913043479	0.6652173913043479	0.6652173913043479	0.6652173913043479
General Medicine	Original	0.018518518518518517	0.018518518518518517	0.018518518518518517	0.018518518518518517
General Medicine	SMOTE	0.5067873303167421	0.5067873303167421	0.5067873303167421	0.5067873303167421
Hematology - Oncology	Original	0.0	0.0	0.0	0.0
Hematology - Oncology	SMOTE	0.8018433179723502	0.8018433179723502	0.8018433179723502	0.8018433179723502
Nephrology	Original	0.0	0.0	0.0	0.0
Nephrology	SMOTE	0.8701923076923077	0.8701923076923077	0.8701923076923077	0.8701923076923077
Neurology	Original	0.043478260869565216	0.043478260869565216	0.043478260869565216	0.043478260869565216
Neurology	SMOTE	0.5603864734299517	0.5603864734299517	0.5603864734299517	0.5603864734299517
Neurosurgery	Original	0.0	0.0	0.0	0.0
Neurosurgery	SMOTE	0.8584474885844748	0.8584474885844748	0.8584474885844748	0.8584474885844748
Obstetrics / Gynecology	Original	0.10344827586206896	0.10344827586206896	0.10344827586206896	0.10344827586206896
Obstetrics / Gynecology	SMOTE	0.7941176470588235	0.7941176470588235	0.7941176470588235	0.7941176470588235
Ophthalmology	Original	0.0	0.0	0.0	0.0
Ophthalmology	SMOTE	0.926605504587156	0.926605504587156	0.926605504587156	0.926605504587156
Orthopedic	Original	0.0	0.0	0.0	0.0
Orthopedic	SMOTE	0.45454545454545453	0.45454545454545453	0.45454545454545453	0.45454545454545453
Pain Management	Original	0.08333333333333333	0.08333333333333333	0.08333333333333333	0.08333333333333333
Pain Management	SMOTE	0.9771689497716894	0.9771689497716894	0.9771689497716894	0.9771689497716894
Pediatrics - Neonatal	Original	0.0	0.0	0.0	0.0
Pediatrics - Neonatal	SMOTE	0.8461538461538461	0.8461538461538461	0.8461538461538461	0.8461538461538461
Psychiatry / Psychology	Original	0.0	0.0	0.0	0.0
Psychiatry / Psychology	SMOTE	0.9223744292237442	0.9223744292237442	0.9223744292237442	0.9223744292237441
Radiology	Original	0.07407407407407407	0.07407407407407407	0.07407407407407407	0.07407407407407407
Radiology	SMOTE	0.4482758620689655	0.4482758620689655	0.4482758620689655	0.4482758620689655
SOAP / Chart / Progress Notes	Original	0.034482758620689655	0.034482758620689655	0.034482758620689655	0.034482758620689655
SOAP / Chart / Progress Notes	SMOTE	0.7688442211055276	0.7688442211055276	0.7688442211055276	0.7688442211055276
Surgery	Original	0.21338912133891214	0.21338912133891214	0.21338912133891214	0.21338912133891214
Surgery	SMOTE	0.05238095238095238	0.05238095238095238	0.05238095238095238	0.05238095238095238
Urology	Original	0.0	0.0	0.0	0.0
Urology	SMOTE	0.7913043478260869	0.7913043478260869	0.7913043478260869	0.791304347826087

When I look at the machine learning metrics of the original data, the values are so low that I do not even need to look at the smote process.



The results are quite poor. Let's apply some domain knowledge and improve the results. The surgery category is kind of a superset because cardiology, neurology, etc. There may be surgeries that require expertise. Similarly, other categories such as Emergency Department Reports, Discharge Summary, and Notes also overlap. That's why I remove them.

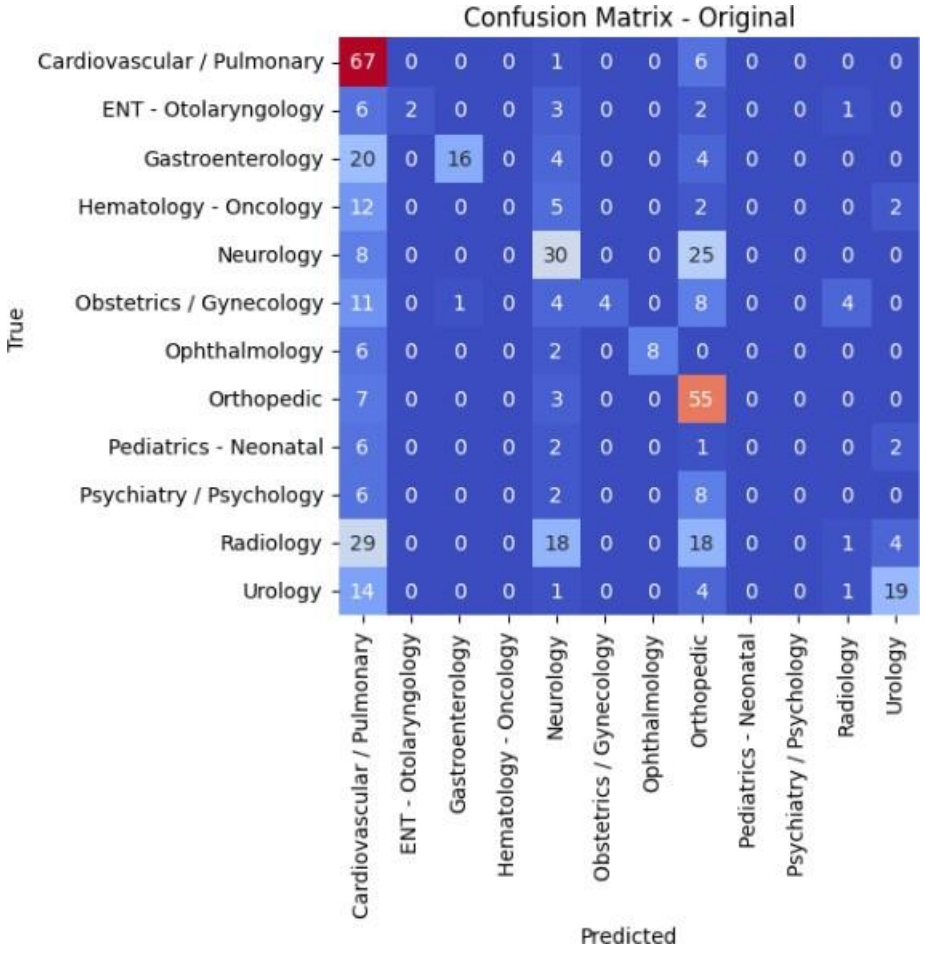


Two-dimensional representation of the matrix obtained after combining sub-medical fields into upper medical fields

4. Application of algorithms for classification

a) Multinomial Naïve Bayes Classification Method

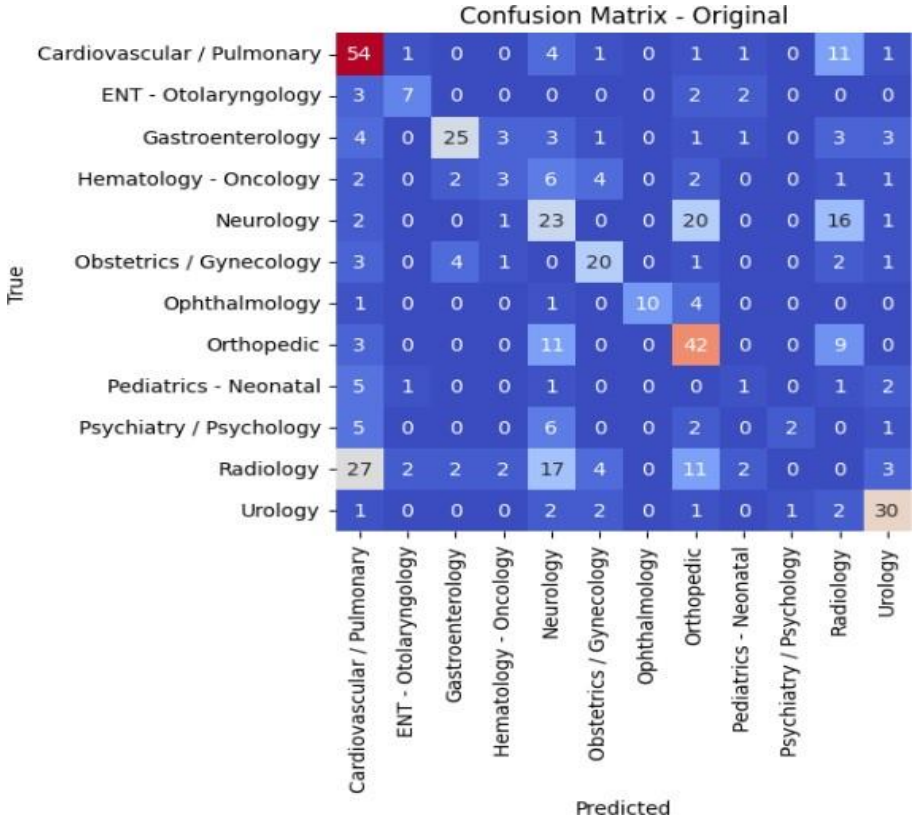
Category	Sampler	Accuracy	Precision	Recall	F1 Score
Cardiovascular / Pulmonary	Original	0.9054054054054054	1.0	0.9054054054054054	0.9503546099290782
Cardiovascular / Pulmonary	SMOTE	0.5540540540540541	1.0	0.5540540540540541	0.7130434782608697
ENT - Otolaryngology	Original	0.14285714285714285	1.0	0.14285714285714285	0.25
ENT - Otolaryngology	SMOTE	0.9090909090909091	1.0	0.9090909090909091	0.9523809523809523
Gastroenterology	Original	0.36363636363636365	1.0	0.36363636363636365	0.5333333333333333
Gastroenterology	SMOTE	0.6710526315789473	1.0	0.6710526315789473	0.8031496062992126
Hematology - Oncology	Original	0.0	1.0	0.0	0.0
Hematology - Oncology	SMOTE	0.8354430379746836	1.0	0.8354430379746836	0.910344827586207
Neurology	Original	0.47619047619047616	1.0	0.47619047619047616	0.6451612903225806
Neurology	SMOTE	0.5757575757575758	1.0	0.5757575757575758	0.7307692307692308
Obstetrics / Gynecology	Original	0.125	1.0	0.125	0.2222222222222222
Obstetrics / Gynecology	SMOTE	0.7702702702702703	1.0	0.7702702702702703	0.870229007633588
Ophthalmology	Original	0.5	1.0	0.5	0.6666666666666666
Ophthalmology	SMOTE	0.9454545454545454	1.0	0.9454545454545454	0.9719626168224299
Orthopedic	Original	0.8461538461538461	1.0	0.8461538461538461	0.9166666666666666
Orthopedic	SMOTE	0.7215189873417721	1.0	0.7215189873417721	0.8382352941176471
Pediatrics - Neonatal	Original	0.0	1.0	0.0	0.0
Pediatrics - Neonatal	SMOTE	0.7605633802816901	1.0	0.7605633802816901	0.864
Psychiatry / Psychology	Original	0.0	1.0	0.0	0.0
Psychiatry / Psychology	SMOTE	0.8767123287671232	1.0	0.8767123287671232	0.9343065693430658
Radiology	Original	0.014285714285714285	1.0	0.014285714285714285	0.028169014084507043
Radiology	SMOTE	0.21839080459770116	1.0	0.21839080459770116	0.3584905660377358
Urology	Original	0.48717948717948717	1.0	0.48717948717948717	0.6551724137931034
Urology	SMOTE	0.625	1.0	0.625	0.7692307692307693



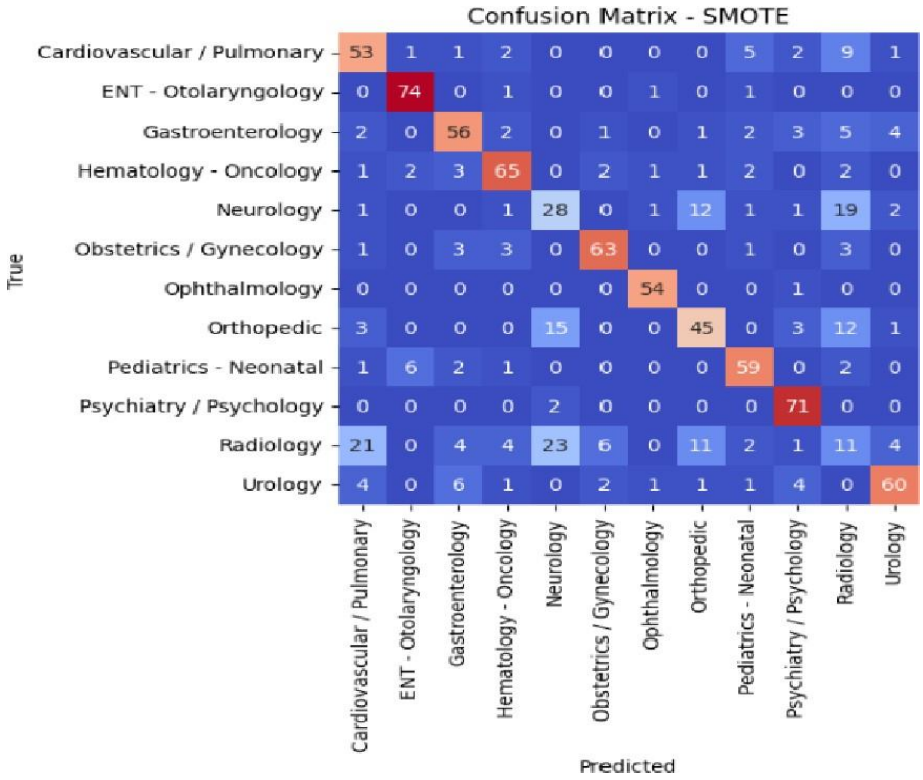
Both the fact that I had to adapt the data I obtained to this classification type (it does not accept values loIr than 0) and the fact that the precision value gives the value 1.0 for each class caused us to eliminate this classification type.(usually indicates that the model fits the training data very Ill, but may have poor generalization ability. The model may be too focused on patterns in the training data and may not be resilient to changes in new data)

b) Random Forest Classification Method

Category	Sampler	Accuracy	Precision	Recall	F1 Score
Cardiovascular / Pulmonary	Original	0.7297297297297297	0.7297297297297297	0.7297297297297297	0.7297297297297297
Cardiovascular / Pulmonary	SMOTE	0.7162162162162162	0.7162162162162162	0.7162162162162162	0.7162162162162162
ENT - Otolaryngology	Original	0.5	0.5	0.5	0.5
ENT - Otolaryngology	SMOTE	0.961038961038961	0.961038961038961	0.961038961038961	0.961038961038961
Gastroenterology	Original	0.5681818181818182	0.5681818181818182	0.5681818181818182	0.5681818181818182
Gastroenterology	SMOTE	0.7368421052631579	0.7368421052631579	0.7368421052631579	0.7368421052631579
Hematology - Oncology	Original	0.14285714285714285	0.14285714285714285	0.14285714285714285	0.14285714285714285
Hematology - Oncology	SMOTE	0.8227848101265823	0.8227848101265823	0.8227848101265823	0.8227848101265823
Neurology	Original	0.36507936507936506	0.36507936507936506	0.36507936507936506	0.36507936507936506
Neurology	SMOTE	0.42424242424242425	0.42424242424242425	0.42424242424242425	0.4242424242424243
Obstetrics / Gynecology	Original	0.625	0.625	0.625	0.625
Obstetrics / Gynecology	SMOTE	0.8513513513513513	0.8513513513513513	0.8513513513513513	0.8513513513513514
Ophthalmology	Original	0.625	0.625	0.625	0.625
Ophthalmology	SMOTE	0.9818181818181818	0.9818181818181818	0.9818181818181818	0.9818181818181818
Orthopedic	Original	0.6461538461538462	0.6461538461538462	0.6461538461538462	0.6461538461538462
Orthopedic	SMOTE	0.569620253164557	0.569620253164557	0.569620253164557	0.569620253164557
Pediatrics - Neonatal	Original	0.09090909090909091	0.09090909090909091	0.09090909090909091	0.09090909090909091
Pediatrics - Neonatal	SMOTE	0.8309859154929577	0.8309859154929577	0.8309859154929577	0.8309859154929577
Psychiatry / Psychology	Original	0.125	0.125	0.125	0.125
Psychiatry / Psychology	SMOTE	0.9726027397260274	0.9726027397260274	0.9726027397260274	0.9726027397260274
Radiology	Original	0.0	0.0	0.0	0.0
Radiology	SMOTE	0.12643678160919541	0.12643678160919541	0.12643678160919541	0.12643678160919541
Urology	Original	0.7692307692307693	0.7692307692307693	0.7692307692307693	0.7692307692307693
Urology	SMOTE	0.75	0.75	0.75	0.75

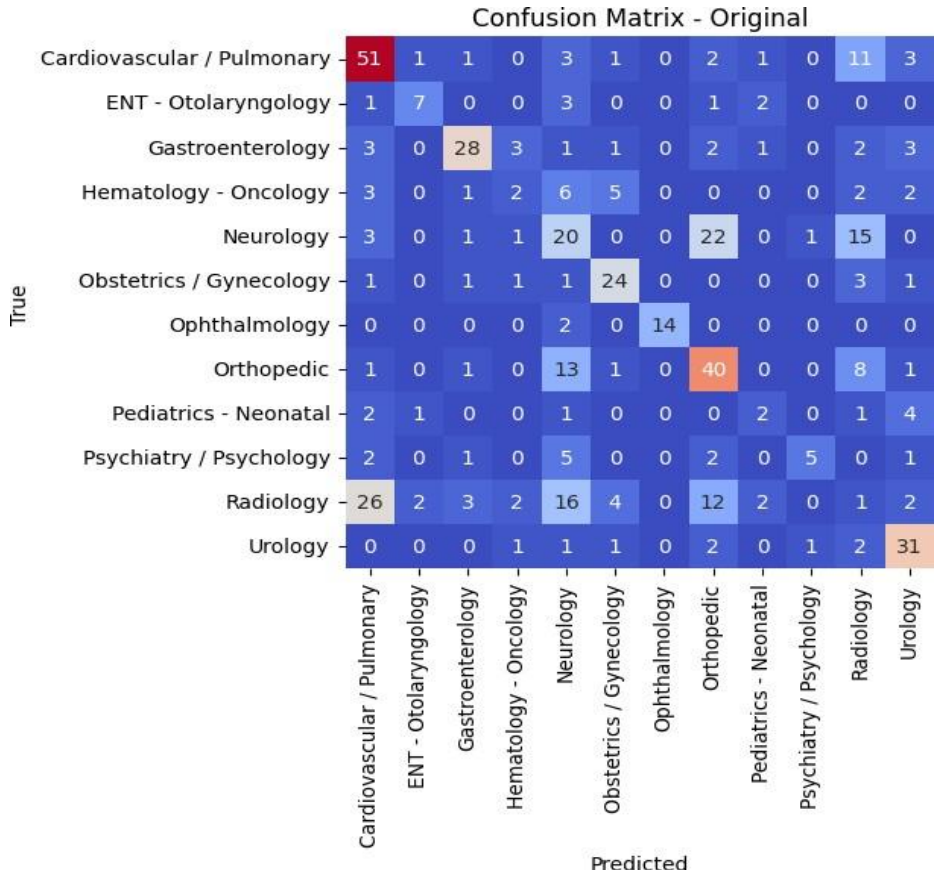


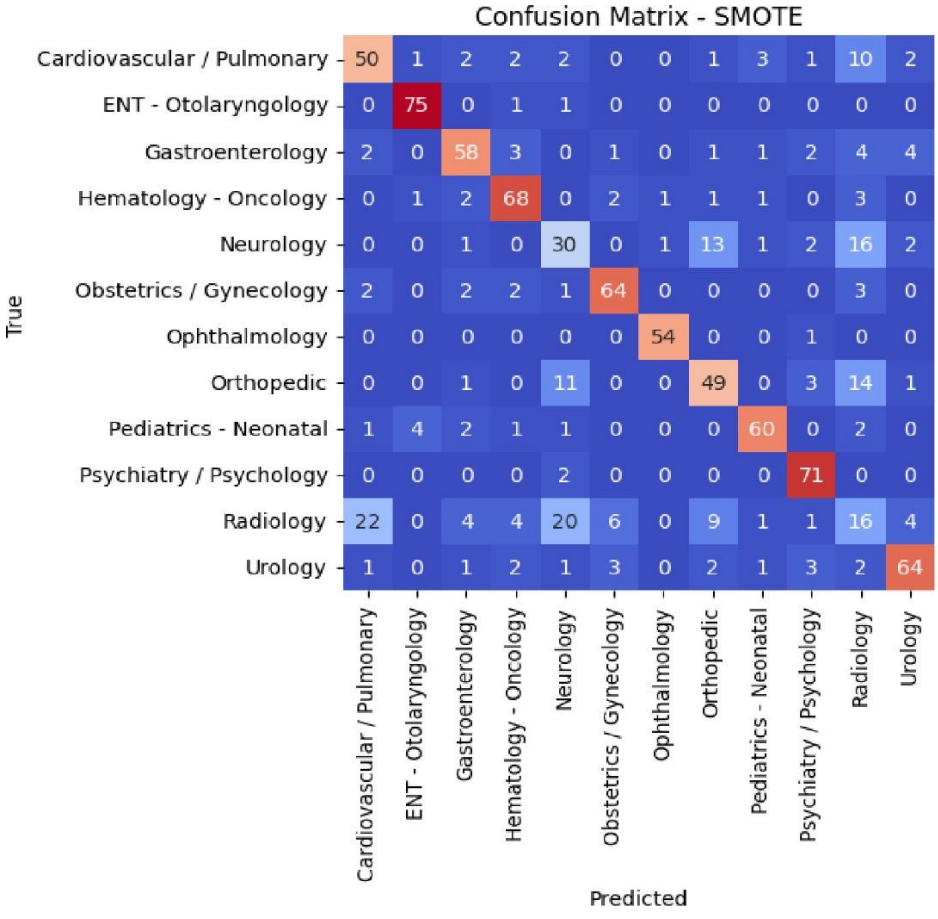
It has been observed that when the Random Forest Classification method is applied, the metrics obtained for each class are very consistent, and when the smote balancing method is applied to the data, there is no information loss as a result of the metrics obtained. We accept the Random Forest Classification Method as successful.



c) Xgboost Classification Method

Category	Sampler	Accuracy	Precision	Recall	F1 Score
Cardiovascular / Pulmonary	Original	0.6891891891891891	0.6891891891891891	0.6891891891891891	0.6891891891891891
Cardiovascular / Pulmonary	SMOTE	0.6756756756756757	0.6756756756756757	0.6756756756756757	0.6756756756756757
ENT - Otolaryngology	Original	0.5	0.5	0.5	0.5
ENT - Otolaryngology	SMOTE	0.974025974025974	0.974025974025974	0.974025974025974	0.974025974025974
Gastroenterology	Original	0.6363636363636364	0.6363636363636364	0.6363636363636364	0.6363636363636364
Gastroenterology	SMOTE	0.7631578947368421	0.7631578947368421	0.7631578947368421	0.7631578947368421
Hematology - Oncology	Original	0.09523809523809523	0.09523809523809523	0.09523809523809523	0.09523809523809523
Hematology - Oncology	SMOTE	0.8607594936708861	0.8607594936708861	0.8607594936708861	0.8607594936708861
Neurology	Original	0.31746031746031744	0.31746031746031744	0.31746031746031744	0.31746031746031744
Neurology	SMOTE	0.45454545454545453	0.45454545454545453	0.45454545454545453	0.45454545454545453
Obstetrics / Gynecology	Original	0.75	0.75	0.75	0.75
Obstetrics / Gynecology	SMOTE	0.8648648648648649	0.8648648648648649	0.8648648648648649	0.8648648648648649
Ophthalmology	Original	0.875	0.875	0.875	0.875
Ophthalmology	SMOTE	0.9818181818181818	0.9818181818181818	0.9818181818181818	0.9818181818181818
Orthopedic	Original	0.6153846153846154	0.6153846153846154	0.6153846153846154	0.6153846153846154
Orthopedic	SMOTE	0.620253164556962	0.620253164556962	0.620253164556962	0.620253164556962
Pediatrics - Neonatal	Original	0.18181818181818182	0.18181818181818182	0.18181818181818182	0.18181818181818182
Pediatrics - Neonatal	SMOTE	0.8450704225352113	0.8450704225352113	0.8450704225352113	0.8450704225352113
Psychiatry / Psychology	Original	0.3125	0.3125	0.3125	0.3125
Psychiatry / Psychology	SMOTE	0.9726027397260274	0.9726027397260274	0.9726027397260274	0.9726027397260274
Radiology	Original	0.014285714285714285	0.014285714285714285	0.014285714285714285	0.014285714285714285
Radiology	SMOTE	0.1839080459770115	0.1839080459770115	0.1839080459770115	0.1839080459770115
Urology	Original	0.7948717948717948	0.7948717948717948	0.7948717948717948	0.7948717948717948
Urology	SMOTE	0.8	0.8	0.8	0.8000000000000002

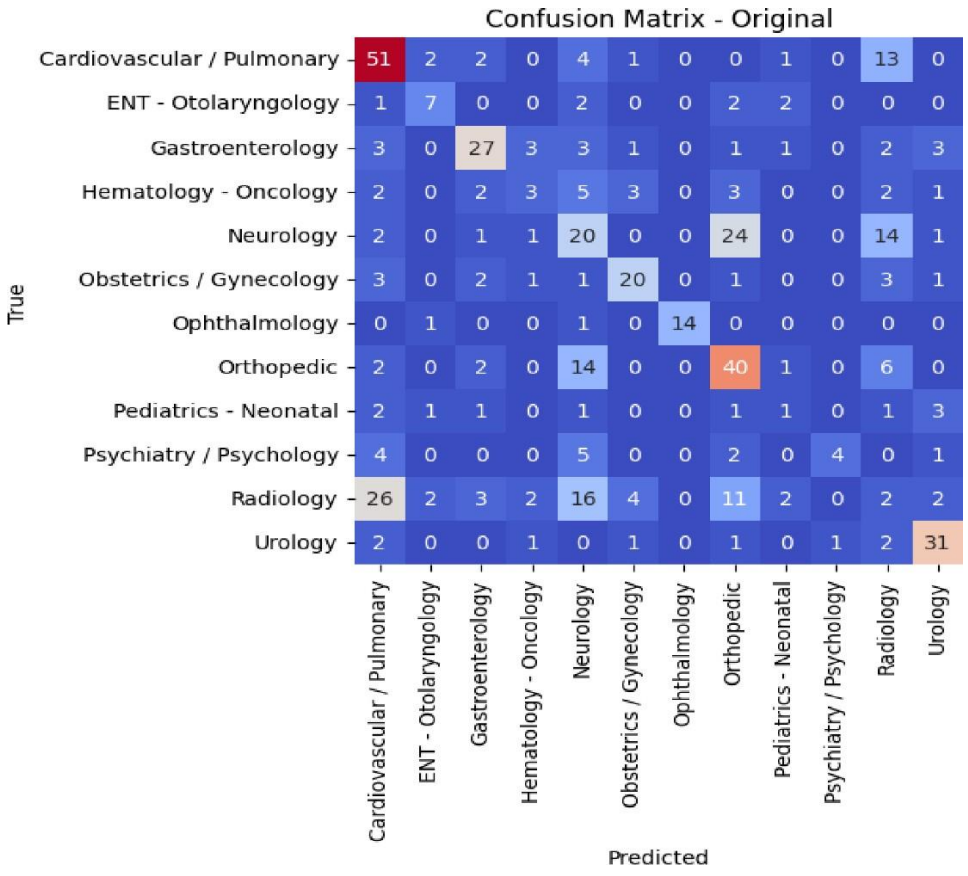

Confusion Matrix - Original
Cardiovascular / Pulmonary 51 1 1 0 3 1 0 2 1 0 11 3
ENT - Otolaryngology 1 7 0 0 3 0 0 1 2 0 0 0
Gastroenterology 3 0 28 3 1 1 0 2 1 0 2 3
Hematology - Oncology 3 0 1 2 6 5 0 0 0 0 2 2
Neurology 3 0 1 1 20 0 0 22 0 1 15 0
Obstetrics / Gynecology 1 0 1 1 1 24 0 0 0 0 3 1
Ophthalmology 0 0 0 0 2 0 14 0 0 0 0 0
Orthopedic 1 0 1 0 13 1 0 40 0 0 8 1
Pediatrics - Neonatal 2 1 0 0 1 0 0 0 2 0 1 4
Psychiatry / Psychology 2 0 1 0 5 0 0 2 0 5 0 1
Radiology 26 2 3 2 16 4 0 12 2 0 1 2
Urology 0 0 0 1 1 1 0 2 0 1 2 31
True
Cardiovascular / Pulmonary
ENT - Otolaryngology
Gastroenterology
Hematology - Oncology
Neurology
Obstetrics / Gynecology
Ophthalmology
Orthopedic
Pediatrics - Neonatal
Psychiatry / Psychology
Radiology
Urology
Predicted

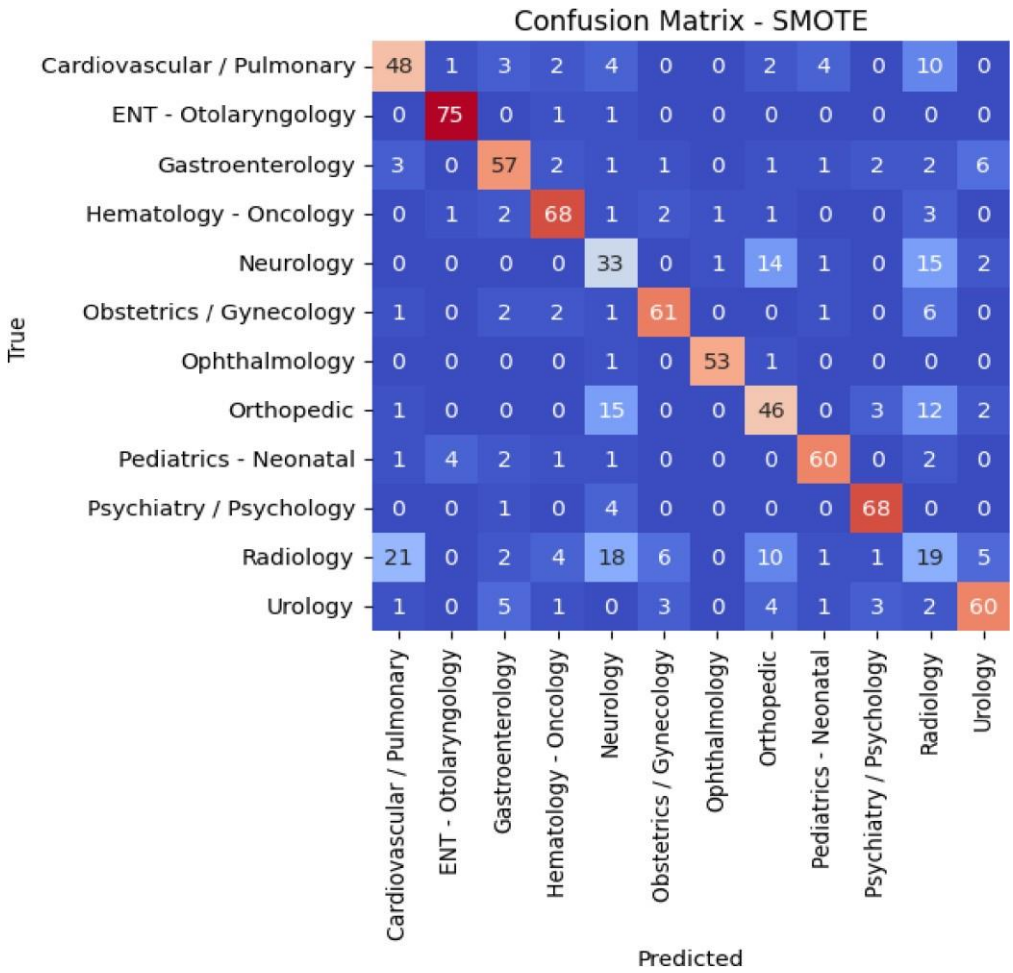

Confusion Matrix - SMOTE
Cardiovascular / Pulmonary 50 1 2 2 2 0 0 1 3 1 10 2
ENT - Otolaryngology 0 75 0 1 1 0 0 0 0 0 0 0
Gastroenterology 2 0 58 3 0 1 0 1 1 2 4 4
Hematology - Oncology 0 1 2 68 0 2 1 1 1 0 3 0
Neurology 0 0 1 0 30 0 1 13 1 2 16 2
Obstetrics / Gynecology 2 0 2 2 1 64 0 0 0 0 3 0
Ophthalmology 0 0 0 0 0 0 54 0 0 1 0 0
Orthopedic 0 0 1 0 11 0 0 49 0 3 14 1
Pediatrics - Neonatal 1 4 2 1 1 0 0 0 60 0 2 0
Psychiatry / Psychology 0 0 0 0 2 0 0 0 0 71 0 0
Radiology 22 0 4 4 20 6 0 9 1 1 16 4
Urology 1 0 1 2 1 3 0 2 1 3 2 64
True
Cardiovascular / Pulmonary
ENT - Otolaryngology
Gastroenterology
Hematology - Oncology
Neurology
Obstetrics / Gynecology
Ophthalmology
Orthopedic
Pediatrics - Neonatal
Psychiatry / Psychology
Radiology
Urology
Predicted

After stating that it is successful in the xgboost classification algorithm, after examining the smote confusion matrix a little, it turns out that it is slightly more successful than the random forest classification method.

d) Lightgbtm Classification Method

Category	Sampler	Accuracy	Precision	Recall	F1 Score
Cardiovascular / Pulmonary	Original	0.6891891891891891	0.6891891891891891	0.6891891891891891	0.6891891891891891
Cardiovascular / Pulmonary	SMOTE	0.6486486486486487	0.6486486486486487	0.6486486486486487	0.6486486486486487
ENT - Otolaryngology	Original	0.5	0.5	0.5	0.5
ENT - Otolaryngology	SMOTE	0.974025974025974	0.974025974025974	0.974025974025974	0.974025974025974
Gastroenterology	Original	0.6136363636363636	0.6136363636363636	0.6136363636363636	0.6136363636363636
Gastroenterology	SMOTE	0.75	0.75	0.75	0.75
Hematology - Oncology	Original	0.14285714285714285	0.14285714285714285	0.14285714285714285	0.14285714285714285
Hematology - Oncology	SMOTE	0.8607594936708861	0.8607594936708861	0.8607594936708861	0.8607594936708861
Neurology	Original	0.31746031746031744	0.31746031746031744	0.31746031746031744	0.31746031746031744
Neurology	SMOTE	0.5	0.5	0.5	0.5
Obstetrics / Gynecology	Original	0.625	0.625	0.625	0.625
Obstetrics / Gynecology	SMOTE	0.8243243243243243	0.8243243243243243	0.8243243243243243	0.8243243243243243
Ophthalmology	Original	0.875	0.875	0.875	0.875
Ophthalmology	SMOTE	0.9636363636363636	0.9636363636363636	0.9636363636363636	0.9636363636363636
Orthopedic	Original	0.6153846153846154	0.6153846153846154	0.6153846153846154	0.6153846153846154
Orthopedic	SMOTE	0.5822784810126582	0.5822784810126582	0.5822784810126582	0.5822784810126582
Pediatrics - Neonatal	Original	0.09090909090909091	0.09090909090909091	0.09090909090909091	0.09090909090909091
Pediatrics - Neonatal	SMOTE	0.8450704225352113	0.8450704225352113	0.8450704225352113	0.8450704225352113
Psychiatry / Psychology	Original	0.25	0.25	0.25	0.25
Psychiatry / Psychology	SMOTE	0.9315068493150684	0.9315068493150684	0.9315068493150684	0.9315068493150684
Radiology	Original	0.02857142857142857	0.02857142857142857	0.02857142857142857	0.02857142857142857
Radiology	SMOTE	0.21839080459770116	0.21839080459770116	0.21839080459770116	0.21839080459770116
Urology	Original	0.7948717948717948	0.7948717948717948	0.7948717948717948	0.7948717948717948
Urology	SMOTE	0.75	0.75	0.75	0.75



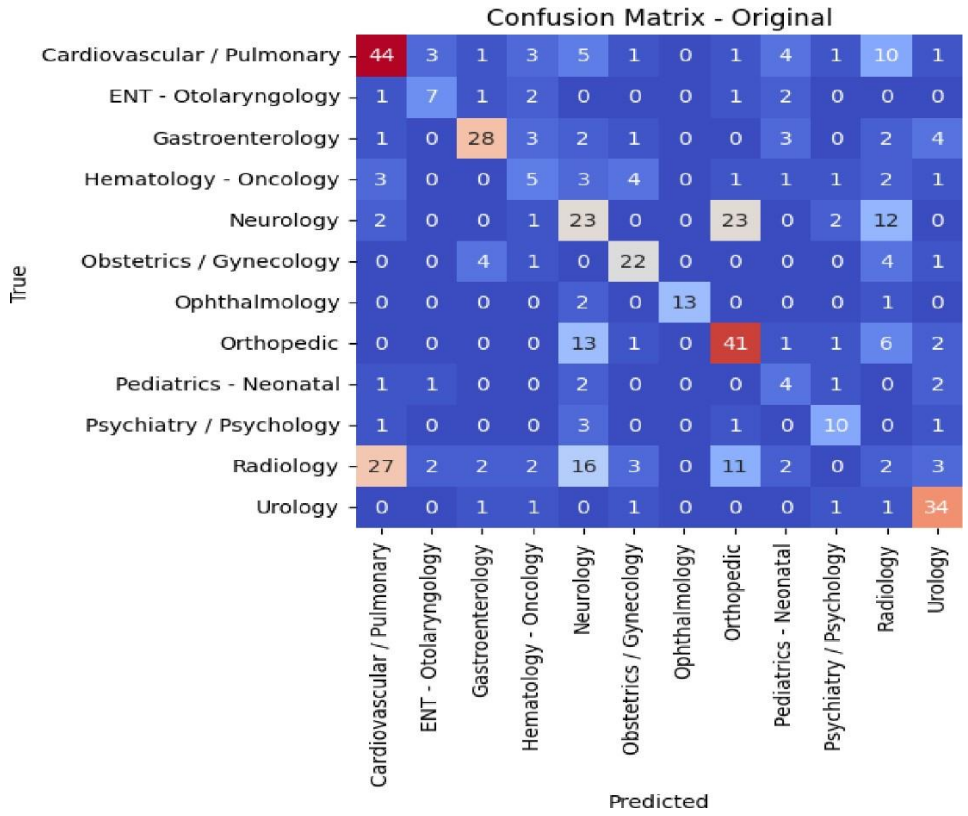


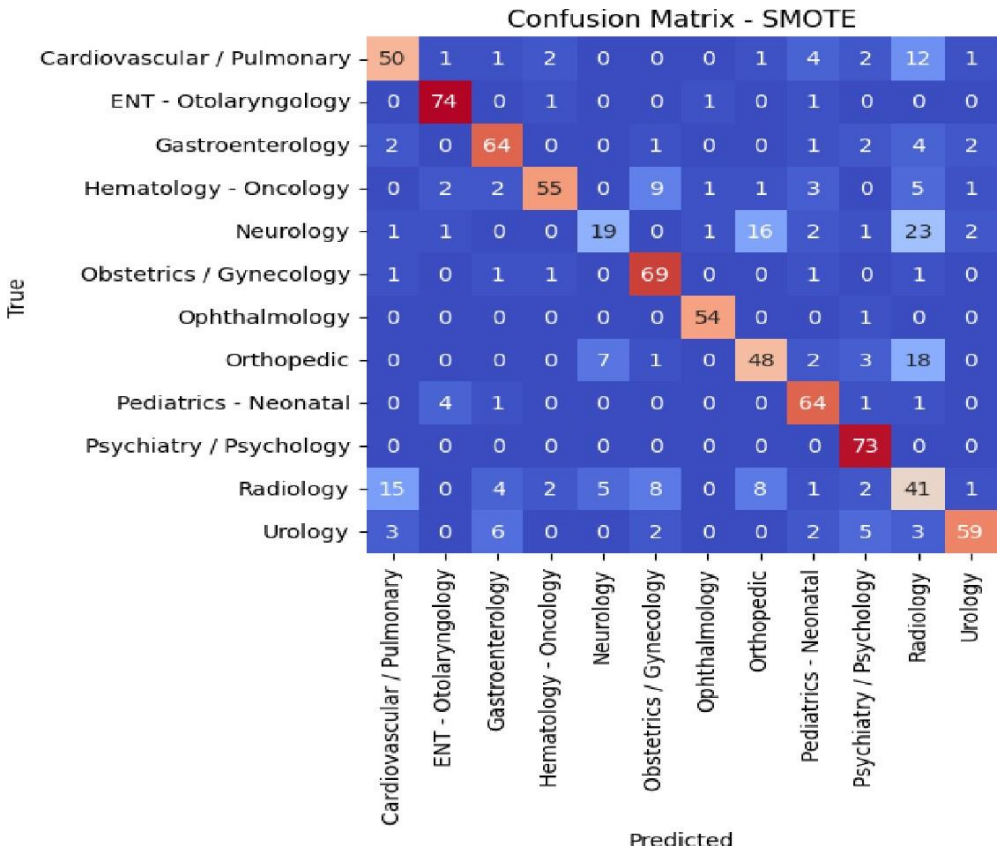
Although there is no informatics loss in the Lightgbm classification method and the metrics are not too bad, the xggbost classification method is more successful.

e) Complex Deep Neural Network Architecture (Ensemble Learning) Using 1D CNN, LSTM

An ensemble model including 1D Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) models was used. The CNN model has 64 filters, 3 kernel sizes and ReLU activation function. This is folloId by a maximum pooling layer of 2. The LSTM model consists of a single 64-unit LSTM layer. By combining the outputs of both models, dense layers are added for classification. Finally, the ensemble model compiled with the Adam optimizer and sparse categorical crossentropy loss is
returned. This model aims to improve overall prediction performance by combining CNN and LSTM architectures.

Category	Sampler	Accuracy	Precision	Recall	F1 Score
Cardiovascular / Pulmonary	Original	0.8378378378378378	0.8378378378378378	0.8378378378378378	0.8378378378378378
Cardiovascular / Pulmonary	SMOTE	0.7027027027027027	0.7027027027027027	0.7027027027027027	0.7027027027027027
ENT - Otolaryngology	Original	0.5	0.5	0.5	0.5
ENT - Otolaryngology	SMOTE	0.948051948051948	0.948051948051948	0.948051948051948	0.948051948051948
Gastroenterology	Original	0.6590909090909091	0.6590909090909091	0.6590909090909091	0.6590909090909091
Gastroenterology	SMOTE	0.8421052631578947	0.8421052631578947	0.8421052631578947	0.8421052631578947
Hematology - Oncology	Original	0.3333333333333333	0.3333333333333333	0.3333333333333333	0.3333333333333333
Hematology - Oncology	SMOTE	0.8481012658227848	0.8481012658227848	0.8481012658227848	0.8481012658227848
Neurology	Original	0.2222222222222222	0.2222222222222222	0.2222222222222222	0.2222222222222222
Neurology	SMOTE	0.4696969696969697	0.4696969696969697	0.4696969696969697	0.4696969696969697
Obstetrics / Gynecology	Original	0.65625	0.65625	0.65625	0.65625
Obstetrics / Gynecology	SMOTE	0.9054054054054054	0.9054054054054054	0.9054054054054054	0.9054054054054054
Ophthalmology	Original	0.6875	0.6875	0.6875	0.6875
Ophthalmology	SMOTE	0.9818181818181818	0.9818181818181818	0.9818181818181818	0.9818181818181818
Orthopedic	Original	0.6461538461538462	0.6461538461538462	0.6461538461538462	0.6461538461538462
Orthopedic	SMOTE	0.6835443037974683	0.6835443037974683	0.6835443037974683	0.6835443037974683
Pediatrics - Neonatal	Original	0.36363636363636365	0.36363636363636365	0.36363636363636365	0.36363636363636365
Pediatrics - Neonatal	SMOTE	0.9577464788732394	0.9577464788732394	0.9577464788732394	0.9577464788732394
Psychiatry / Psychology	Original	0.375	0.375	0.375	0.375
Psychiatry / Psychology	SMOTE	1.0	1.0	1.0	1.0
Radiology	Original	0.02857142857142857	0.02857142857142857	0.02857142857142857	0.02857142857142857
Radiology	SMOTE	0.08045977011494253	0.08045977011494253	0.08045977011494253	0.08045977011494253
Urology	Original	0.8205128205128205	0.8205128205128205	0.8205128205128205	0.8205128205128205
Urology	SMOTE	0.8	0.8	0.8	0.8000000000000002





While I can say that the metric results obtained in the neural network model I applied are successful, there is a problem arising from the field of radiology medicine, as in other classification algorithms.

5. Conclusion

In our project, I tried to classify medical specialties accurately according to transcript texts. In this direction, four different traditional machine learning classification algorithms and one CNN + LSTM ensemble learning algorithm Ire applied. As a result of the experiments, it was observed that especially xgboost and ensemble learning classification methods Ire successful.

The success of xgboost and ensemble learning methods shows that these models are effective in correctly classifying health-related texts. This can allow for more effective management of medical documents and transcripts and rapid access to healthcare professionals.

Holver, it should be emphasized that these results can be further strengthened with more data and more studies. Further work on a large data set to evaluate the consistency of the classification methods used in the project may provide a more

comprehensive evaluation of model performance. In this way, it can be established on a more solid basis which classification method is more consistent and how reliable the project is in general.

As a result, it was concluded that this study shows that xgboost and ensemble learning classification methods used to accurately classify medical specialties are successful and that these successes can be strengthened with further studies. This is a promising step towards effective management of medical documents and faster access to healthcare professionals.

References

https://www.kaggle.com/code/ritheshsreenivasan/clinical-text-classification

https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions/data