Feature extraction using openSMILE feature

namely “ComParE_2016.conf” for openly available pathological speech dataset-2

Pathological speech dataset II

- Speech databases recorded primarily for medical research but is useful in linguistics as well
- Two databases were recorded:
  - for healthy children's speech (recorded in kindergarten and in the first level of elementary school)
  - for pathological speech of children with a Specific Language Impairment (recorded at a surgery of speech and language therapists and at the hospital).

Development Dysphasia

- Also known as Specific Language Impairment a language disorder that **delays the mastery of language skills** in children who have **no hearing loss or other developmental delays.**
- These children fail to acquire their native language properly/completely, despite having normal non-verbal intelligence, no hearing problems, and no known neurological dysfunctions or behavioral, emotional or social problems
- It is estimated that SLI affects approximately 5–7% of the kindergarten population

The Database

The entire database contains three subgroups of recordings of children's speech from different types of speakers.

- The first subgroup (controls, or H-CH) consists of recordings of children without speech disorders;
- the second subgroup (cases, or SLI-CH I) consists of recordings of children with SLI;
- the third subgroup (cases, or SLI-CH II) consists of recordings of children who have SLI of different degrees of severity (1 –mild, 2 –moderate, or 3 –severe).

Trends in the way that speech changed during the given time period (approximately 3 months) were the determining factors for inclusion in SLI-CH I database, rather than the degree of severity of the children's diagnosis.

Table 2. Description of All Databases. Subgroup H-CH is for controls and subgroups SLI-CH I and SLI-CH II are for cases.

	H-CH		SLI-CH I	SLI-CH II
		WITH DEFECT		
Girls	45	16	13	26
Number of recordings	45	16	22	45
Boys	25	17	33	46
Number of recordings	25	17	64	88
All children	70	33	46	72
All recordings	70	33	86	133
All utterances	4620	2178	5676	8819

Table 1. Speech database–structure and types of utterances used in our research.

Task code	Type of part	# Patterns	Descripton	
[T 1]	Vowels	5	Czech	"a", "o", "u", "e", "i"
			English	"a", "o", "u", "e", "i"
[T 2]	Consonants	10	Czech	"m", "b", "t", "d", "r", "l", "k", "g", "h", "ch"
			English	"m", "b", "t", "d", "r", "l", "k", "g", "h", "ch"
[T 3]	Syllables	9	Czech	"pe", "la", "vla", "pro", "bě", "nos", "ber", "krk", "prst"
			English	"pe", "la", "vla", "for", "bě", "nose", "take", "neck", "finger"
[T 4]	Two-syllable words	5	Czech	"kolo", "pivo", "sokol", "papír", "trdlo"
			English	"wheel", "beer", "falcon", "paper", "boob"
[T 5]	Three-syllable words	4	Czech	"dědeček", "pohádka", "pokémon", "květina"
			English	"grandfather", "fairy tale", "Pokemon", "flower"
[T 6]	Four-syllable words	3	Czech	"motovidlo", "televize", "popelnice"
			English	"niddy noddy", "television", "dustbin"
[T 7]	Difficult words	2	Czech	"r znobarevný", "mateřídouška"
			English	"varicoloured", "thyme"
[T 8]	Geminate words	3	Czech	"pohádková víla", "kouzelný měšec", "čarotvorný hrnec"
			English	"fairy", "magic pouch", "magic pot"
[T 9]	Accretion of range of words	4	Czech	"voda", "živá voda", "živá a mrtvá voda", "pramen s živou a mrtvou vodou"
			English	"water", "live water", "live and dead water", "source of live and dead water"
[T 10]	Sentence	1	Czech	"Když šla červená Karkulka k babičce, potkala zlého vlka."
			English	"When Little Red Riding Hood went to her grandmother, she met bad wolf."
[T 11]	Auditory differentiation	10	Czech	"pes—nes", "ten—den", "k l—v l", "hrát—brát", "ječí—ježí", "ble—ple", "kloč—kloč", "kvěš—kveš", "šný—šní", "vošl—vočl"
			English	Change in one phoneme in the word. For example: "pes—nes", . . .
[T 13]	Describe the picture	1	English	"Look at the laughable clown."—A spontaneous description of the girl's picture.

OpenSMILE

- The Munich **open toolkit Speech and Music Interpretation by Large Space Extraction** (openSMILE) is a modular and flexible feature extractor for signal processing and machine learning applications.
- It contains a number of default feature sets, some of which are shown:

- Chroma features for key and chord recognition
- MFCC for speech recognition
- PLP for speech recognition
- Prosody (Pitch and loudness)
- The INTERSPEECH 2009 Emotion Challenge feature set
- The INTERSPEECH 2010 Paralinguistic Challenge feature set
- The INTERSPEECH 2011 Speaker State Challenge feature set
- The INTERSPEECH 2012 Speaker Trait Challenge feature set
- The INTERSPEECH 2013 ComParE feature set
- The MediaEval 2012 TUM feature set for violent scenes detection.

OpenSMILE ComParE_2016 feature set

- It has approximately 200 features divided into two groups of Low-Level Descriptors (LLDs).
  - 59 LLDs in group A, 54 functionals are applied
  - 59 delta LLDs of group A, 46 functionals are applied
  - 6 LLDs in group B, 39 functionals are applied
  - 6 delta LLDs of set B, 39 functionals are applied.
- This results in a total of 6,368 features.

LLD and Functionals

Table 3.5 INTERSPEECH 2013 Computational Paralinguistics ChallengE (Com- ParE) set

LLD	Functionals
Group A: (59)	Arithmetic^{A*,B} or positive arithmetic$^{A\delta,B}$ mean,
Loudness,	Root-quadratic mean, flatness,
Modulation loudness,	Standard deviation, skewness, kurtosis,
RMS energy, ZCR,	Quartiles 1–3,
RASTA auditory bands 1–26,	Inter-quartile ranges 1–2, 2–3, 1–3,
MFCC 1–14,	99th and 1-st percentile, range of these,
Energy 250–650 Hz,	Relative position of max. and min. value,
Energy 1–4 kHz,	Range (maximum to minimum value),
Spectral RoP .25, .50, .75, .90,	Linear regression slope^{A*,B} a, offset^{A*,B} b,
Spectral flux, entropy, variance,	Linear regression quadratic error^{A*,B},
Spectral skewness and kurtosis,	Quadratic regression coeff.^{A*,B} a, b, c,
Spectral slope,	Quadratic regression quadratic error^{A*,B},
Spectral harmonicity,	Temporal centroid^{A*,B},
Spectral sharpness (auditory),	Peak mean value^{A} and dist. to arithm. mean^{A},
Spectral centroid (linear).	Mean^{A} and std. dev.^{A} of peak to peak distances,
Group B: (6)	Peak and valley range^{A} (absolute and relative),
	Peak-valley-peak slopes mean^{A} and std. dev.^{A},
F_0 via SHS, Prob. of voicing,	Segment length mean^{A}, min.^{A}, max.^{A}, std. dev.^{A},
Jitter (local and delta),	Up-level time 25 %, 50 %, 75 %, 90 %,
Shimmer,	Rise time, left curvature time,
logḢNR (time domain).	Linear Prediction gain and coefficients 1–5.

Overview of Low-level Descriptors (LLDs) and functionals applied to these LLDs. Functionals marked with ^{A} and ^{B} are only applied to group A or B LLDs (and deltas), respectively; functionals marked with * or $^{\delta}$ are not or only (respectively) applied to the delta LLDs. Details in Appendix A.1.5

Work Completed Till Mid Evaluation

1. Downloaded and understood the dataset and its divisions.
2. Set up openSMILE
3. Extracted the ComParE_2016 feature set for utterances in our dataset.
4. Worked on an automated script for extracting ComParE_2016 feature set for all .wav files in the dataset.
5. [In progress] Writing the code to extract another feature from the dataset.

Work Completed After Mid Eval

1. Extracted ComParE_2016 feature set for all healthy speaker utterances and patient utterances in the dataset
2. Trained 3 types of classifiers - random forest, SVM and K-means with 5-fold cross validation
3. Extracted MFCC features for all healthy speaker utterances and patient utterances in the dataset
4. Trained the same set of models for this feature set - cross validation and train_test
5. Compared and contrasted the results given by the two feature sets

Classification using ComParE feature

Method: Cross-Validation
Total number of samples: 3250

We saw the that the best performance was given by the SVM model. We also noticed that the accuracy for determining pathological speech is consistently greater.

Results

SVM

Metric	Healthy	Patient
Recall	0.9308755760368663	0.9792626728110599
Precision	0.957345971563981	0.9659090909090909
F1-Score	0.9439252336448597	0.9725400457665904

SVM

Metric	Healthy	Patient
Recall	0.9631336405529954	0.9654377880184332
Precision	0.9330357142857143	0.9812646370023419
F1-Score	0.9478458049886621	0.9732868757259

SVM

Metric	Healthy	Patient
Recall	0.9212962962962963	0.9746543778801844
Precision	0.9476190476190476	0.9613636363636363
F1-Score	0.9342723004694835	0.9679633867276888

SVM

Metric	Healthy	Patient
Recall	0.9212962962962963	0.9722863741339491
Precision	0.943127962085308	0.9611872146118722
F1-Score	0.9320843091334895	0.9667049368541907

SVM

Metric	Healthy	Patient
Recall	0.9444444444444444	0.976905311778291
Precision	0.9532710280373832	0.9724137931034482
F1-Score	0.9488372093023255	0.9746543778801843

kMeans

Metric	Healthy	Patient
Recall	0.12442396313364056	0.6728110599078341
Precision	0.15976331360946747	0.6058091286307054
F1-Score	0.13989637305699484	0.6375545851528384

kMeans

Metric	Healthy	Patient
Recall	0.9032258064516129	0.31336405529953915
Precision	0.3967611336032389	0.8662420382165605
F1-Score	0.5513361462728551	0.46023688663282575

kMeans

Metric	Healthy	Patient
Recall	0.9120370370370371	0.31336405529953915
Precision	0.397979797979798	0.8774193548387097
F1-Score	0.5541490857946554	0.4617996604414262

kMeans

Metric	Healthy	Patient
Recall	0.8935185185185185	0.3233256351039261
Precision	0.39711934156378603	0.8588957055214724
F1-Score	0.5498575498575499	0.46979865771812085

kMeans

Metric	Healthy	Patient
Recall	0.8842592592592593	0.3394919168591224
Precision	0.40041928721174	0.8546511627906976
F1-Score	0.5512265512265513	0.4859504132231405

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.8847926267281107	0.9654377880184332
Precision	0.927536231884058	0.9436936936936937
F1-Score	0.9056603773584906	0.9544419134396355

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.8894009216589862	0.9723502304147466
Precision	0.9414634146341463	0.9461883408071748
F1-Score	0.9146919431279621	0.9590909090909091

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.9074074074074074	0.9815668202764977
Precision	0.9607843137254902	0.9551569506726457
F1-Score	0.9333333333333333	0.9681818181818181

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.8935185185185185	0.976905311778291
Precision	0.9507389162561576	0.9484304932735426
F1-Score	0.9212410501193318	0.9624573378839589

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.888888888888888	0.976905311778291
Precision	0.9504950495049505	0.9463087248322147
F1-Score	0.9186602870813396	0.9613636363636363

MFCC Feature Extraction

- Mel-Frequency analysis of speech is based on human perception experiments
- It is observed that human ear acts as filter. It concentrates on only certain frequency components
- These filters are non-uniformly spaced on the frequency axis:
  - More filters in the low frequency regions
  - Less no. of filters in high frequency regions
- Filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the mel-frequency scale

MFCC Feature Extraction

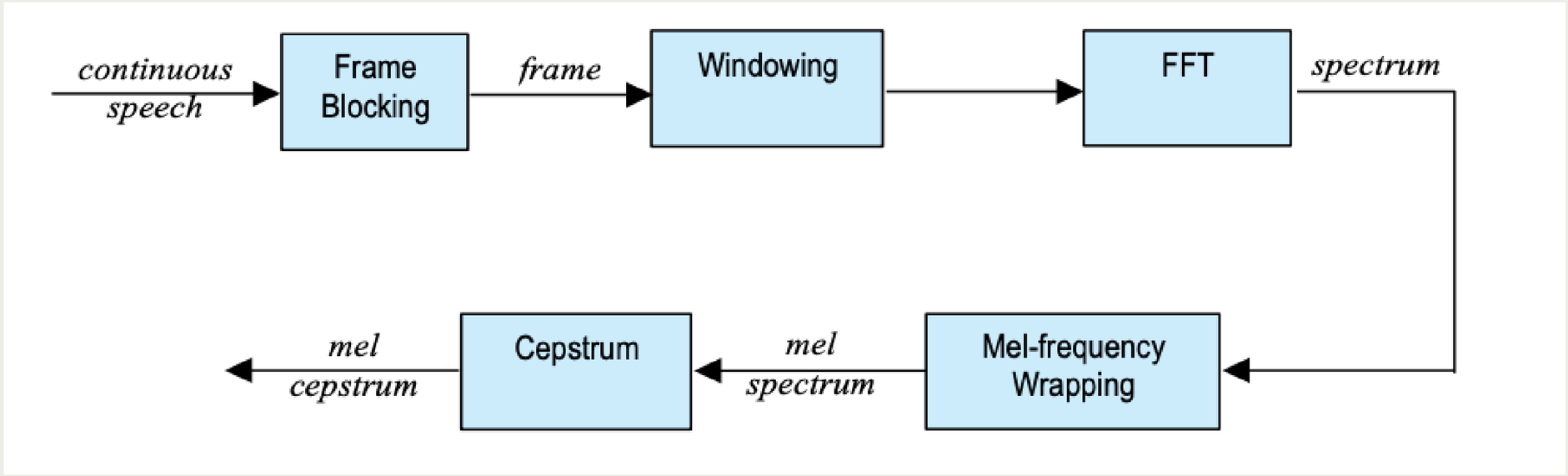
- The relationship between frequency in Hz and frequency in Mel scale is given by:

$$m = 1125 \ln(1 + \frac{f}{700})$$

$$f = 700 \, (e^{\frac{m}{1125}} - 1)$$

MFCC Feature Extraction

Process:



MFCC Feature Extraction

Step 1: Pre-emphasis: The speech signal s(n) is sent to a high-pass filter:

$$y(t) = x(t) - \alpha x(t - 1)$$

where typical values for the filter coefficient (α) are 0.95 or 0.97.

The goal of pre-emphasis is to compensate the high-frequency part that was suppressed during the sound production mechanism of humans. Moreover, it can also amplify the importance of high-frequency formants.

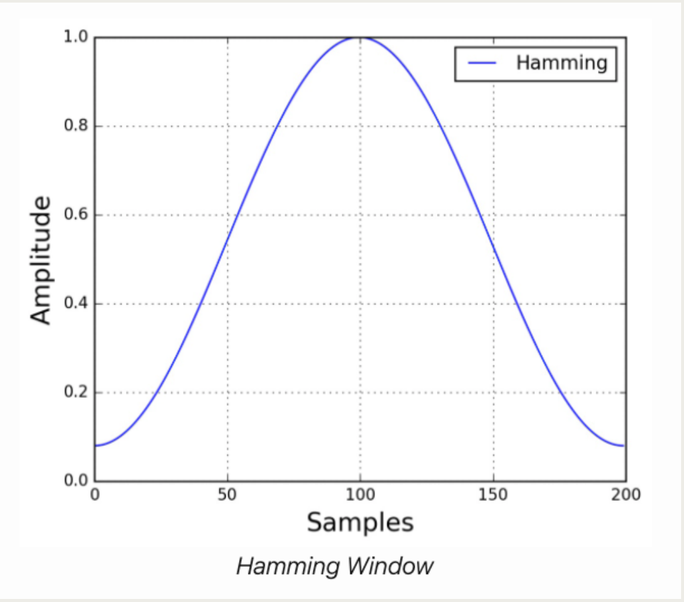
MFCC Feature Extraction

Step 2: Frame Blocking: The input speech signal is segmented into frames of 20-30 ms with optional overlap of 1/3 or 1/2 of the frame size. In our case:

$$\text{frameSize} = 0.020 * \text{fs}$$
$$\text{overlap} = (\text{frameSize}/2)$$

MFCC Feature Extraction

Step 3: Windowing: We multiply each frame by a Hamming window to increase its continuity at the first and last points.



Hamming Window

MFCC Feature Extraction

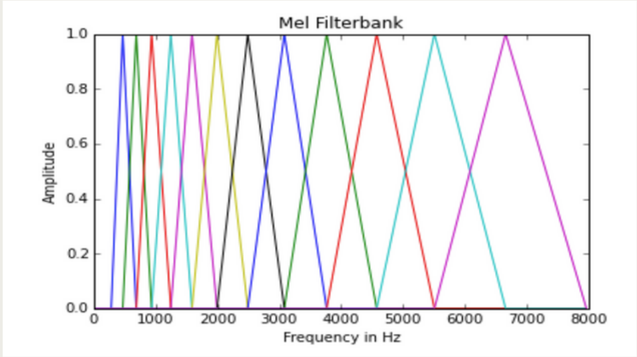
Step 4: FFT and periodogram: After windowing, Fast Fourier Transform (FFT) is applied to find the power spectrum of each frame.

$$P = \frac{|FFT(x_i)|^2}{N}$$

where, xi is the ith frame of signal x.

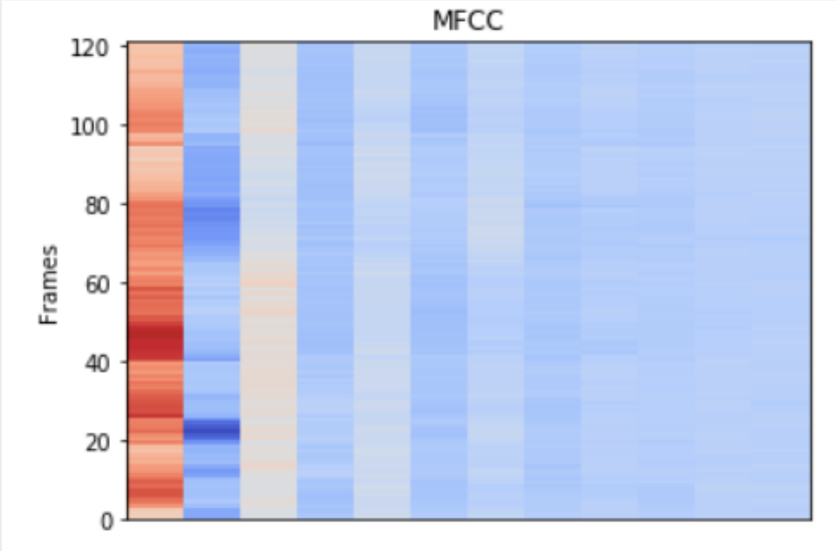
MFCC Feature Extraction

Step 5: Filter Banks: The entire frequency range is divided into ‘n’ Mel filter banks, which is also the number of coefficients we want. To calculate filter bank energies we multiply each filter bank with the power spectrum, and add up the coefficients. Once this is performed we are left with ‘n’ numbers that give us an indication of how much energy was in each filter bank.



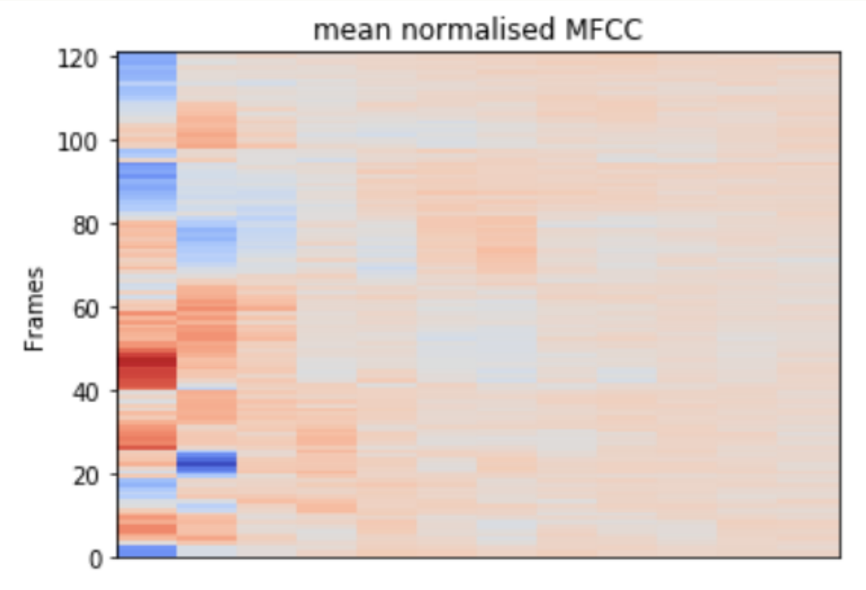
MFCC Feature Extraction

Step 6: DCT: We take the logarithm of these ‘n’ energies and compute its Discrete Cosine Transform to get the final MFCCs.



MFCC Feature Extraction

Mean normalisation: MFCC values are not very robust in the presence of additive noise, and so it is common to normalise their values in speech recognition systems to lessen the influence of noise.



Classification using MFCC

Method: Cross-Validation
Total number of samples: 3250

We saw the that the best performance was given by the SVM model. We also noticed that the accuracy for determining pathological speech is less as compared to when done using ComParE feature.

Results

SVM

Metric	Healthy	Patient
Recall	0.5345622119815668	0.8827586206896552
Precision	0.6946107784431138	0.7917525773195876
F1-Score	0.6041666666666666	0.8347826086956521

SVM

Metric	Healthy	Patient
Recall	0.576036866359447	0.8827586206896552
Precision	0.7102272727272727	0.8067226890756303
F1-Score	0.6361323155216285	0.8430296377607025

SVM

Metric	Healthy	Patient
Recall	0.5622119815668203	0.896551724137931
Precision	0.7305389221556886	0.8041237113402062
F1-Score	0.6354166666666667	0.8478260869565218

SVM

Metric	Healthy	Patient
Recall	0.5	0.8732718894009217
Precision	0.6625766871165644	0.7782340862422998
F1-Score	0.5699208443271767	0.8230184581976113

SVM

Metric	Healthy	Patient
Recall	0.6111111111111112	0.868663594470046
Precision	0.6984126984126984	0.8177874186550976
F1-Score	0.6518518518518519	0.8424581005586592

kMeans

Metric	Healthy	Patient
Recall	0.7596899224806202	0.6309963099630996
Precision	0.494949494949495	0.8465346534653465
F1-Score	0.599388379204893	0.7230443974630021

kMeans

Metric	Healthy	Patient
Recall	0.7751937984496124	0.6125461254612546
Precision	0.4878048780487805	0.8512820512820513
F1-Score	0.5988023952095809	0.7124463519313305

kMeans

Metric	Healthy	Patient
Recall	0.7906976744186046	0.6051660516605166
Precision	0.4880382775119617	0.8586387434554974
F1-Score	0.6035502958579881	0.7099567099567099

kMeans

Metric	Healthy	Patient
Recall	0.689922480620155	0.6273062730627307
Precision	0.46842105263157896	0.8095238095238095
F1-Score	0.5579937304075235	0.7068607068607068

kMeans

Metric	Healthy	Patient
Recall	0.7829457364341085	0.6531365313653137
Precision	0.517948717948718	0.8634146341463415
F1-Score	0.6234567901234569	0.7436974789915966

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.4186046511627907	0.959409594095941
Precision	0.8307692307692308	0.7761194029850746
F1-Score	0.5567010309278351	0.8580858085808581

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.43410852713178294	0.966789667896679
Precision	0.8615384615384616	0.7820895522388059
F1-Score	0.577319587628866	0.8646864686468646

RandomForestClassifier

Metric	Healthy	Patient
Recall	0.3875968992248062	0.955719557195572
Precision	0.8064516129032258	0.7662721893491125
F1-Score	0.5235602094240838	0.8505747126436782

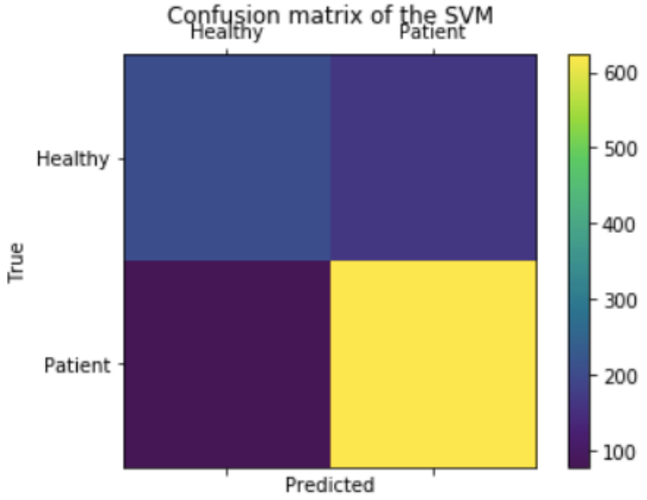
RandomForestClassifier

Metric	Healthy	Patient
Recall	0.4186046511627907	0.966789667896679
Precision	0.8571428571428571	0.7774480712166172
F1-Score	0.5625000000000001	0.8618421052631579

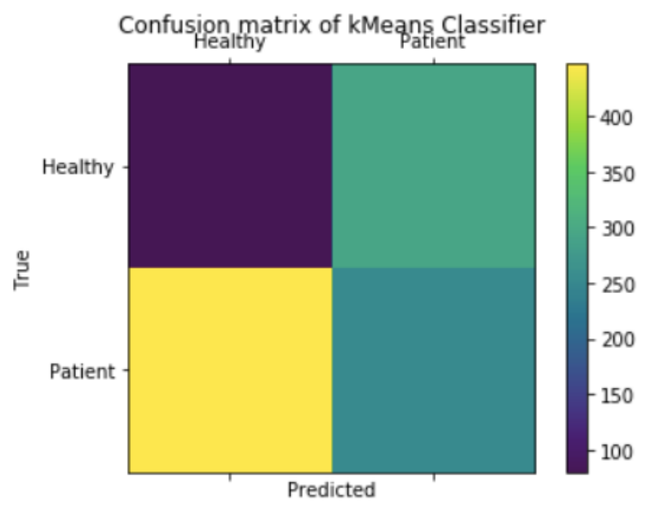
RandomForestClassifier

Metric	Healthy	Patient
Recall	0.4108527131782946	0.915129151291513
Precision	0.6973684210526315	0.7654320987654321
F1-Score	0.5170731707317073	0.8336134453781513

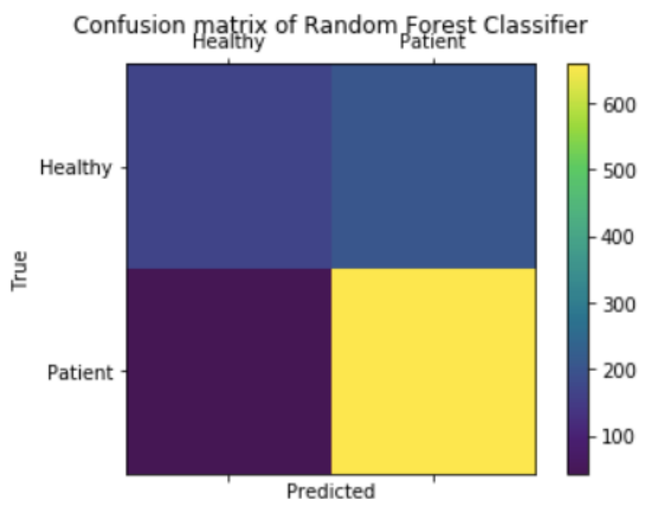
Train_test results



SVM		
Metric	Healthy	Patient
Recall	0.553475935828877	0.8887303851640513
Precision	0.7263157894736842	0.7886075949367088
F1-Score	0.6282245827010622	0.835680751173709
Accuracy	0.772093023255814	



kMeans		
Metric	Healthy	Patient
Recall	0.21390374331550802	0.362339514978602
Precision	0.15180265654648956	0.4635036496350365
F1-Score	0.1775804661487236	0.4067253803042434
Accuracy	0.31069767441860463	



Random Forest		
Metric	Healthy	Patient
Recall	0.44919786096256686	0.9386590584878745
Precision	0.7962085308056872	0.7615740740741
F1-Score	0.5743589743589743	0.8408945686900958
Accuracy	0.7683720930232558	

Thank you!

Presented by
GROUP 8

Harshita Sharma & Aashna Jena