Project title: Predicting Wine Quality

Team Name: D^3A

Team member Name	SRN
Name1: Amiya Mishra	PES2UG19CS034
Name 2: Deboleena Mukherjee	PES2UG19CS102
Name 3: Deepali Suraj Attavar	PES2UG19CS106
Name 4: Deepthi B	PES2UG19CS107

1. Dataset Name and Description.

The Dataset is a wine-quality dataset that is publicly available for research purposes from UCI- Machine Learning. The dataset contains real data on vihn verde from northwest Portugal. The dataset consists of 1599 red samples. Data were collected from May 2004 to February 2007 using iLab computerized system that automatically manages the process of wine samples testing starting from the manufacturer's requirements to laboratory and sensory analysis. The below table presents the physicochemical statistics for the dataset.

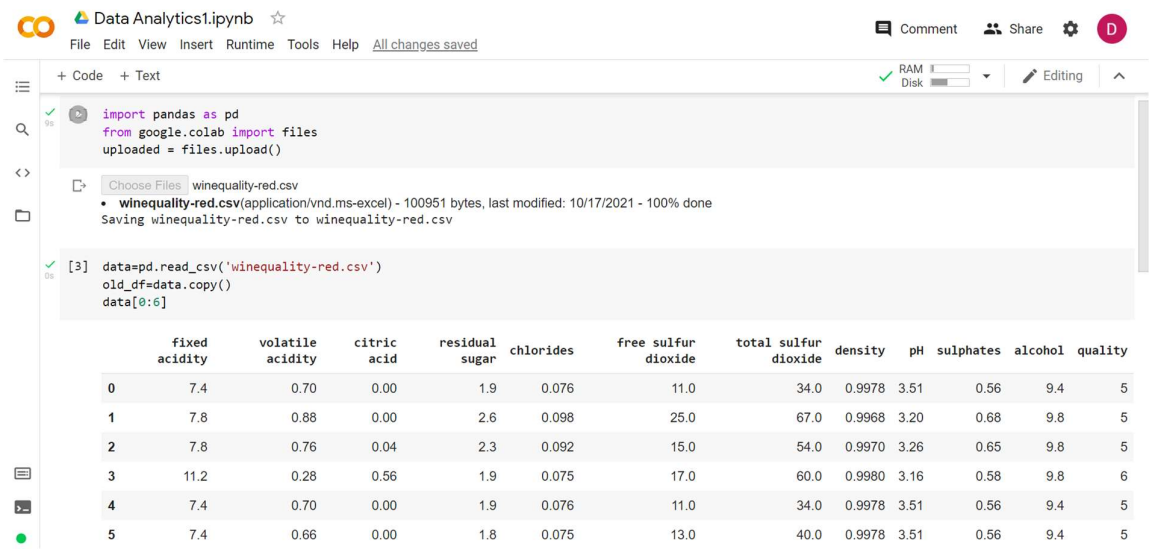
Table 1: The physicochemical data statistics per wine type

	Red wine(1599 instances)	
	Mın	Max
fıxed acıdıty (*g/dm³*)	4.60	15.9
volatıle acıdıty (*g/dm³*)	0.12	1.58
cıtrıc acıd (*g/dm³*)	0.00	1.00
resıdual sugar (*g/dm³*)	0.9	15.5
chlorıdes	0.01	0.61
free sulfur dıoxıde (*mg/dm³*)	1	72
total sulfur dıoxıde (*mg/dm³*)	6	289
densıty (*g/cm³*)	0.99	1.00
ph	2.74	4.01
sulphates (*g/dm³*)	0.33	2.00
alcohol *(% vol.)*	8.4	14.9
qualıty	3.00	8.00

2. Problem statement:

To determine the quality of wine, sensory tests are used which rely on human expert's knowledge, but physicochemical properties of wine can also be used. The relationship between physicochemical and sensory analysis are complex and not yet fully understood, but significant correlations can be found between quality of wine and physicochemical properties. Data mining techniques are powerful techniques to analyze relationships between different attributes of a dataset. They can be used for classification, clustering, forecasting, optimization, and summarization. In wine industry, DM is used to make recommendations on purchase of wine, based on wine ratings, consumer criticisms, and wine prices. There are a large number of websites and mobile applications that make recommendations for choosing wines based on that information (Ex: www.go-wine.com). Despite its potential to predict wine quality based on physicochemical properties, DM techniques aren't often used in this task. In this project, we present a classification of wines based on their physicochemical properties that are easily measurable and accessible. This analysis can be valuable to wine producers (to improve the production process), to consumers (to select wine), and to experts to support their evaluation of wine and to potentially improve the speed and quality of their decisions.

3. EDA and Visualization



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
5	7.4	0.66	0.00	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5

- How many rows and attributes?

No of rows: 1599, No of attributes: 11. One output variable.

- How many missing data and outliers?

```
[4]  data.isnull().sum().sum()

     0
```

No missing data.

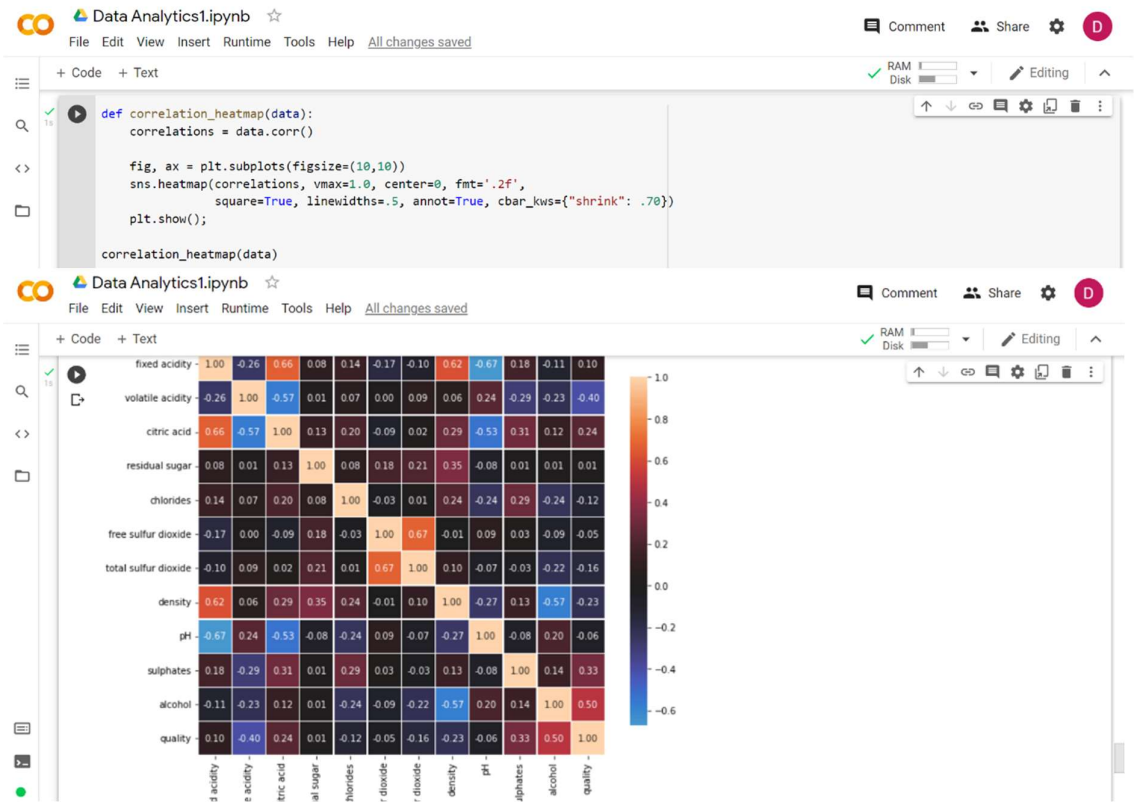
Doing boxplot on all the attributes, we found that the data consists of outliers.

- Any inconsistent, incomplete, duplicate or incorrect data?

No, there is no incomplete, inconsistent, or incorrect data. This is because the data was collected using iLab computerized system which is a sophisticated system that prevents such data entries. And therefore, the data is mostly clean.

- Are the variables correlated to each other?

We decided to visualise a correlation matrix.



The strongly correlated items are:

1.fixed acidity and citric acid.
2.free sulphur dioxide and total sulphur dioxide.
3.fixed acidity and density.
4.alcohol and quality.
So, from above points there is a clear inference that alcohol is the most important characteristic to determine the quality of wine.

The weekly correlated items are:

1.citric acid and volatile acidity.
2.fixed acidity and ph.
3.density and alcohol.

These are some relations which do not depend on each other at all.

- Are any of the pre-processing techniques needed: dimensionality reduction, range transformation,standardization, etc.?
  This doesn't require dimensionality reduction as of now, as the attributes as different classifiers can take different no. of inputs. As there are only 11 attributes, which are all important in deciding the quality. However, as seen in the case studies, attributes can be removed while choosing and applying the model.

4. **Link for google sheet:**

https://docs.google.com/spreadsheets/d/1WM9Cnt9R0nEpPxBKed1Ui8Ch3lloLcG4tKklG6Rz4K0/edit#gid=0

5. **Link for Google Colab:**

https://colab.research.google.com/drive/1KokAX_ZtJyS1ai-1Tb32RezPLAkX7z3z#scrollTo=rIfnX6S5N4bk

6. Literature Survey:

1) TITLE: Classification based on Data-Mining Approach for Quality Control in Wine Prediction

Authors: P.Appalasamy, A.Mustapha, N.D Rizal, F. Johari, A.F Mansor
Publisher, year: Asian Network for Scientific Information 2012
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of model used: Decision Tree: ID3, Naive Bayes
Accuracy: ID3:60%, Naive Bayes: 58.8%
First Data is discretised during Pre-processing where continuous values are converted to discrete nominal data using first-last method. Then Attribute Selection takes place to limit curse of dimensionality- into 4 attributes only. This includes: Volatile acidity, Total sulphur dioxide, Sulphate, Alcohol. This is done by Info Gain Eval method by WEKA software tool. Training data is divided into 10 sets and each algorithm is applied iteratively 10 times.ID3 and NB are the algorithms. Both, when applied on test data give poor performance/accuracy. However, ID3 performs better. NB finds highly correlated attributes for its classification

2) TITLE: Wine Quality prediction Model Using Machine Learning Techniques

Authors: Rohan Dilip Kothawade
Publisher, year: University of Skovde, 2021
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: Support Vector Machine, Naive Bayes, Artificial Neural Network
Accuracy: SVM: 83.52% Naive Bayes:46.33% ANN:85.16%
First, Data that is unbalanced is dealt with using SMOTE technique where unbalanced classes are over-sampled to make it balanced. Then feature selection is done based on correlation analysis- rank is given to the attribute. Standardising is done to fit data between 0 and 1. Splitting data where test size is 0.2. Hyperparameter tuning done to get gamma, kernel, c which help deriving the prediction function for each of the models SVM and ANN. It can be seen that ANN performs best with accuracy of 85.16% on the test data followed by SVM among the models.

3) TITLE: A Data Mining Approach to Wine Quality Prediction

Authors: Dragana Radosavljevic, Sinisa Illic, Stefan Pitulic
Publisher, year: International Scientific Conference, Unitech, Gabrovo, 2019

Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: Decision Tree: J48 Random Forest: RF Naive Bayes: NB Multilayer Perceptron: MLP, Support Vector Machine: SVM K *
Accuracy: Categorization1:(9 labels) J48:60.63% RF:68.33% NB:59.38% MLP:62.92% SVM:57.08% K*:61.67%
Categorization2:(3 labels) J48:82.71% RF:87.29% NB:84.38% MLP:85.21% SVM:81.46% K*:83.96%
In this project many models are used, and for each model a different pre-processing technique provided by the WEKA software tool. There is no algorithm that provides high accuracy. This is called the No Free Lunch Theorem. Model built in 2 ways: -10-fold cross validation method (dividing data into 10 portions) -Using test sample method- training set: test set =2/3:1/3 with test of significance=0.05, using RESAMPLE filter in WEKA software. Hyperparameter Tuning involves calculating Kappa, Mean Absolute Error, Root mean square error and other performance measures like ROC and Recall J48 decision tree algorithm can deal with overfitted data, pruning, rule derivation and so on. MLP is a feed forward neural network that maps random data set to corresponding outputs. Random Forest is a classifier that produces the best result among a cluster of predictors. SVM is based on the simple idea: to define a hyperplane that separates data into appropriate classes. K* algorithm is instance-based classifier that is the class of a test instance is based upon the class of those instances that are similar to it. Based on results, it can be seen that RF gives the best performance.

4) TITLE: Assessing wine quality using a decision tree.

Authors: Seunghan Lee, Juyoung Park, Kyungtae Kang
Publisher, year: 2015 IEEE International Symposium on Systems Engineering (ISSE)
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: Decision tree: ID3
Accuracy: ID3 Accuracy for red wine=58.7%, ID3 Accuracy for red wine=58.7%
This model is been built based on decision tree to infer the quality as perceived by a consumer. First the data is used to summarize the scope of physiochemical data. The table summarizes the importance of each item of the physicochemical data. Then the decision tree is constructed on physiochemical characteristics of the wine and predicts the taste preference. This is done by recursive subdivision and selecting the most influential attribute at every instance of a node using entropy and information gain. It then splits the set of instances into subsets and to become a child node. The ID3 decision tree is complete when all instances are distributed as leaf nodes. The model is compared to weka's implementation of 3 machine learning algorithms(libSVM, BayesNet, MultiPerceptron) of the same dataset. The results show that the model outperforms the others.

5) TITLE: Modelling Wine Quality from physicochemical properties.

Authors: Dale Angus
Publisher, year: Stanford University.
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: Neural network model: binary classifier and multi-class classifier.
Accuracy: Multi class classifier top-2 categorical accuracy for red wine=0.9500 and white wine=0.8888. Binary classifier achieved 98% accuracy.
First a quick survey was performed in scikit learn to study the feature selection and to determine the best prediction performance. Further the focus was on development and understanding of neural networks using various modulus and software packages. 2 models were developed: multi-class classifier and a binary classifier. For the multi-class classifier, the data was divided into train, validation, and test in ratio: 56-24-20. An experiment on different aspects such as regularization, number of layers and various activation functions. Based on the result of the experiment, the optimizer was found and studied to conclude. For the binary classifier the classifier model had to predict whether the wine was red wine or white wine. The accuracy achieved was 98% for 2 features combinations and 99% accuracy for 4 features combinations.

6) TITLE: Final Report SENG474: Wine Analysis

Authors: Noah Spriggs, Murray Dunne, Chris Life, Greg Richardson, Haoyan Xu
Publisher, year: University of Victoria
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: Naive bayes, SVM, Prism algorithm, decision tree.
Accuracy: For this paper, rmse was used as validation metric. Naive bayes rmse for red wine=0.7706 and rmse for white wine=0.9455. SVM rmse for red wine=0.772971 and white wine=0.837515. Prism algorithm rmse for red wine=0.7226 and white wine=0.7983. Decision tree rmse for red wine=0.8249 and for white wine=0.7577
This paper report approach began with choosing rmse as the validation metric to unify the comparative tool. The model was test against Gaussian Naive bayes to decide whether multinomial was worth pursuing. But due to poor results, this classification is suitable for text involved than continuous numeric data. A python script for SVM (AVA, OVA) implementation on wine quality data. The AVA and OVA results were same for the data. Using the Prism algorithm to model the dataset, the result compared favourably to other models used and more accurate than Navie Bayes, and SVM but less accurate than linear regression and random decision tree. The decision tree model was then built and gave the best results out of all other classifiers. After combining all results, comparing rmse, rectifying the errors, re-doing, random

decision trees gave the lowest rmse and is the best model to predict wine quality dataset.

7) TITLE: Wine Quality Prediction Using Data Mining

Authors: Shruthi P
Publisher, year: ATME college of Engineering
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: neural network and SVM
Accuracy: Naive Bayes=(for 36) Accuracy 100%, Simple Logistic Classifier(178)=Accuracy 97.22%, KStar Classifier(178)=Accuracy 97.22%, JRip Classifier(178)=Accuracy 94.44%
Different classification algorithms (Naive Bayes, Simple Logistic, KStar, JRip, J48) are applied on same data set of 178 wine samples, all the five algorithm's efficiency is good but Naive Bayes is more accurate classifier among all. These algorithms can be used to classify the wine to the respective quality levels. It helps the consumers of wine and reduces the number of fraud in wine industry. It also helps the quality labelling companies of government to issue quality certificate. It reduces the errors compared to manual quality assurance. In future cost of wine can also be predicted based on the quality assessment done using data mining techniques. Other classification and prediction algorithms can also be applied and studied in future.

8) TITLE: Red Wine Quality Prediction Using Machine Learning Techniques

Authors: Sunny Kumar, Kanika Agarwal, Nelshan Mandan
Publisher, year: IIT Roorkee
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: Naive Bayes, Decision Tree and SVM
Accuracy: Navi Bayes =55.89% SVM= 68.64%
Data mining nowadays is most important technique which is utilized for investigation of the archives. It looks at the information and produces the required yield. With the headway in the innovation, it helps in playing the sound test in the market thus benefits the client. As a result of its property of investigating the information it is utilized in the examination to process diverse execution appraisals utilizing different calculations. In this exploration accuracy, precision, misclassification error, F-score, recall and specificity are resolved. Since the training dataset contains about 70% of the data from the original dataset, thus the results demonstrate the Support Vector Machine as

the best algorithm giving an accuracy of 67.25% implemented on red wine quality prediction on RStudio software, then comes Random Forest giving an accuracy 65.83% and last comes the Naïve Bayes algorithm giving an accuracy of 55.91%. In future, better algorithms can be developed which involves the combination of best features of all other data mining techniques. If certain adjustments in the hyperplane, and balanced tree technique along with the appropriate probability are used then much better

9) TITLE: Prediction of quality of wines using Supervised ML learnings

Authors: Satyabrata Aich, Ahmed Al-Absi, Mangal Sain
Publisher, year: Inje University
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model: Random Forest, SVM
Accuracy: Random Forest= 70.33%, SVM= 66.54%
Each classifier able to show all the performance metrics such as accuracy, sensitivity, specificity, PPV, and NPV based on the test data. We have applied all the classification techniques to the GA based reduced feature sets for two types of wine as well as SA based reduced feature sets for two types of wine to measures the performance parameter with respect to each classifier. We separated each performance measures with respect to GA and SA sets and plot the column plot for better visualization.

10) TITLE: Selection of important features and predicting wine quality using machine learning techniques.

Authors: Yogesh Gupta
Publisher, year: GLA University
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model: Neural network, SVM, linear regression
This paper explores the usage of machine learning techniques such as linear regression, neural network and support vector machine for product quality in two ways. Firstly, determine the dependency of target variable on independent variables and secondly, predicting the value of target variable. In this paper, linear regression is used to determine the dependency of target variable on independent variables. On the basis of computed dependency, important variables are selected those make significant impact on dependent variable. Further, neural network and support vector machine are used to predict the values of dependent variable. All the experiments are performed on Red Wine datasets. This paper proves that better prediction can be made if selected features (variables) are being considered rather than considering all the features. Therefore, this paper explores different machine learning

techniques such as linear regression, neural networks (NN) and support vector machines (SVM) for product quality assurance. These techniques perform quality assurance process with the help of available characteristics of product and automate the process by minimizing human interfere. The work also identifies the important features to predict the values of dependent variables.

11) TITLE: A new red wine prediction framework using machine learning

Authors: Gupta, Appalaswamy, Mustapha, Abedin, Cortez
Publisher, year: College of physical and Electronics engineering, Sichuan Normal University, Chengdu China
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model used: Multiple regression, SVM, Neural networks
Accuracy: Multiple regression-59.1%, SVM-59.1%, Neural Network(NN)-62.4%, ID3-60%, Naive Bayesian-58.8%, k-NN-71.9% (selected features)
Correct understanding of red wine physicochemical properties is the basis and premise for red wine quality success. We proposed a new framework combined MF-DCCA with XGBoost and LightGBM. For the correlation importance and classification results that is obtained, we think that the proposed approach is an advance in red wine quality classification. Residual sugar contributes the most complexly to red wine quality while the weakest cross-correlation are volatile acidity and chlorides, respectively. Both LightGBM and XGBoost achieved higher classification accuracy than the other machine learning algorithms. The relative importance of the inputs brought significant views regarding the impact of the analytical tests. Since some physicochemical variables can be controlled in the production process this information can be used to improve the red wine quality. The proposed new framework is based on objective tests and thus it can be integrated into a decision support system, improving the speed and quality of the oenologist performance. Hence, we expect this will help to promote red wine in research and development processes.

12) TITLE: The role of machine learning in productivity: A case study of Wine quality prediction

Publisher, year: European Journal of science and technology
Dataset used: winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009
Name of the model: SVM, decision tree, clustering techniques, WSVM, SMOTE
Productivity can be defined as a measure of the efficiency of a person, machine, production systems, etc., in converting inputs into useful outputs. As

different from a few decades ago, autonomous systems have become more popular and have been used in our age to increase the productivity of a system. As computers become capable of learning freely, reasoning, and determining the best course-of-action in real-time, they are started to be integrated into the real production systems to increase productivity. When it is said learning ability, the first things that come to the mind is artificial intelligence (AI) and machine learning (ML). Machine learning systems that take place under the umbrella of artificial intelligence provide splendid learning capabilities to computers. There available numerous machine learning approaches such as Multilayer Perceptrons, Support Vector Machines, Decision Trees, Clustering methods used in a broad spectrum of domains. The main algorithm chosen to utilize is Support Vector Machines which is a commonly used machine learning method in various applications. In addition to using baseline model SVM, we have applied weighting data samples and producing synthetic samples strategies. This study has focused on the binary classification problem (desired and undesired products). However, it can be easily extended to a multiclass classification.

7. Your Plan

Our plan was to collect different research paper and collect data according to that and find accuracy and error according to that and finally select best of them. And find consequences regarding of our all the datas, we can predict the quality of wines easily seeing our report.

8. References

winequality-red.csv (winequality-white.csv, not taken in my analysis) Cortez et al, 2009

Wikipedia:- for finding reference about wines distribution in the country and best quality present in different states.

Literature survey paper links:
https://ui.adsabs.harvard.edu/abs/2012JApSc..12..598A/abstract
https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1574730&dswid=-4036
https://unitech-selectedpapers.tugab.bg/images/papers/2019/s5/s5_p120.pdf
https://ieeexplore.ieee.org/abstract/document/7302752/
http://cs229.stanford.edu/proj2019aut/data/assignment_308832_raw/25895690.pdf
https://webdocs.cs.ualberta.ca/~alona/dm_projs/wine_final_report.pdf
https://ieeexplore.ieee.org/abstract/document/9063846
https://ieeexplore.ieee.org/abstract/document/9104095
https://www.sciencedirect.com/science/article/pii/S1877050917328053