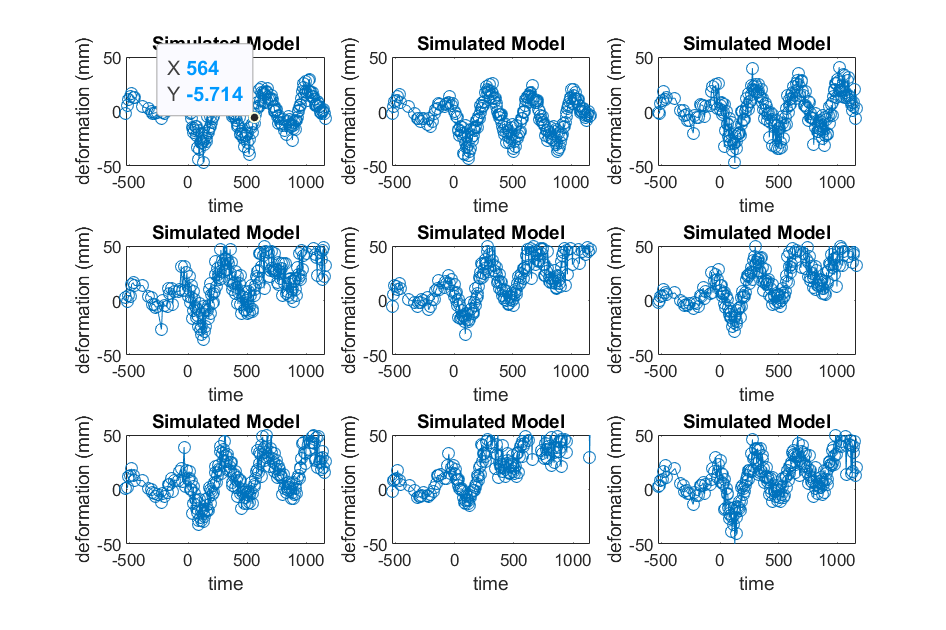
**Proiect Statistical Learning**

1. **ÎNVĂŢARE SUPERVIZATĂ: CLASIFICARE**

În fişierul **Training\_Set\_Labeled\_data.txt** sunt stocate, sub formă matriceală, pe coloane, 1000 de date etichetate cu 0 sau 1 (câte 500 din fiecare clasă) . Acest set de date va reprezenta setul de antrenare al modelului de machine learning. O coloană are pe ultima poziţie eticheta corespunzătoare, restul valorilor reprezentând valorile datelor. Din punct de vedere fizic, fiecare dată reprezintă un profil de deformare. Mai jos este un exemplu a 9 astfel de date. Dimensiunea fiecărei date este 219.



În fişierul **Test\_Set\_Labeled\_data.txt** sunt stocate 200 de date (tot pe coloane, în acelaşi format ca in cazul fişierului **Training\_Set\_Labeled\_data.txt)** , fişierul conţinând datele de test pentru modelele de clasificare.

**Analiza clasificării în spatiul datelor**

Folosind datele etichetate să se realizeze o clasificare folosind metodele:

1. REGRESIA LOGISTICĂ (RL)
2. LDA
3. QDA
4. KNN
5. SVM
6. RANDOM FOREST
7. ADABOOST

În antrenarea clasificatoarelor se vor folosi exclusiv datele din Training Set. După calibrarea fiecărei metode, aceasta se va testa cu datele din setul de testare

Se cere să se reprezinte performanţa fiecărui clasificator (Matricea de confuzie, acuratetea, precizia, regasirea, Scor F1). Performanţa metricilor se va prezenta comparativ pe cele doua seturi de date (training si test)

**Analiza clasificarii în spaţiul redus dimensonal**

Să se concateneze cele două seturi de date, formându-se un nou set de 1200 de profile. Să se realizeze o reducere dimensională folosind PCA, mapând datele într-un spaţiu bi-dimensional. Să se prezinte maparea. Folosind datele etichetate din spatiul bi-dimensional să se realizeze o clasificare folosind metodele:

1. REGRESIA LOGISTICĂ (RL)
2. LDA
3. QDA
4. KNN
5. SVM
6. RANDOM FOREST
7. ADABOOST

În antrenarea clasificatoarelor se vor folosi exclusiv datele din Training set (obs: ale proiecţiilor lor cu PCA in spaţiul bi-dimensional). După calibrarea fiecărei metode, aceasta se va testa cu datele din setul de testare (obs. tot proiectiile).

1. Se cere să se reprezinte performanţa fiecărui clasificator (Matricea de confuzie, acuratetea, precizia, regasirea, Scor F1). Performanţa se va prezenta comparativ pe cele doua seturi de date (training si test).
2. Pentru clasificatorii RL, LDA, QDA, SVM să se reprezinte grafic limita de detecţie împreună cu datele din ambele seturi (de testare şi antrenare, colorate diferit).

**Analiza comparativă a clasificatorilor**

Să se realizeze un tabel în care sa se prezinte diferentele intre metrici pentru clasificatorii antrenati in spatiul datelor respectiv in spatial redus dimensional.

1. **ÎNVĂŢARE NESUPERVIZATĂ: REDUCERE DIMENSIONALĂ ŞI CLUSTERING**

În fişierul **Data\_Red\_Dim.txt** se găsesc sub formă matriceală, pe coloane, 2000 de date etichetate cu 0, 1,2,3 (câte 500 din fiecare clasă). Datele aparţin a 4 profile de deformare diferite.

1. Să se realizeze o reducere dimensională printr-o proiecţie într-un spaţiu bi-dimensional a acestor date folosind metodele:
2. MDS
3. PCA
4. LDA
5. t-SNE
6. ISOMAP
7. UMAP

**Cerintă: Unde se preteaza, in cazul modelelor de reducere dimensională, se cere realizarea unui fine tunning. Utilizarea valorilor default pentru hiperparametri nu se punctează maxim. Punctajul maxim se obtine pentru fine tuning.**

Obs: În prezentarea mapărilor să se prezinte şi etichetele prin colorare diferită

1. Pentru fiecare reducere dimensională să se realizeze o clusterizare folosind tehnicile
2. GMM
3. DBSCAN
4. HIERACHICAL CLUSTERING
5. KMEANS

**Cerintă: Se cere realizarea unui fine tuning. Utilizarea valorilor default pentru hiperparametri nu se punctează maxim. Punctajul maxim se obtine pentru fine tuning.**