Stochastic Processes
CS4070

Quarter 1 Final Assignment

Name: **Konstantinos Krachtopoulos**
Student Number: **5472539**

Contents

Table of Figures

Table of Tables

Exercise 1

Exercise 1.1

In order to estimate the autocorrelation, we need to have the distribution of the signal. In practice, we only have different realizations of a signal.

Exercise 1.2

The 10 random realizations can be seen in the graph below. As a cross-check, we notice that the points of the realizations fall around the value 0, and rarely deviate more than 2 ($2 \cdot \sigma^2$) units from zero.

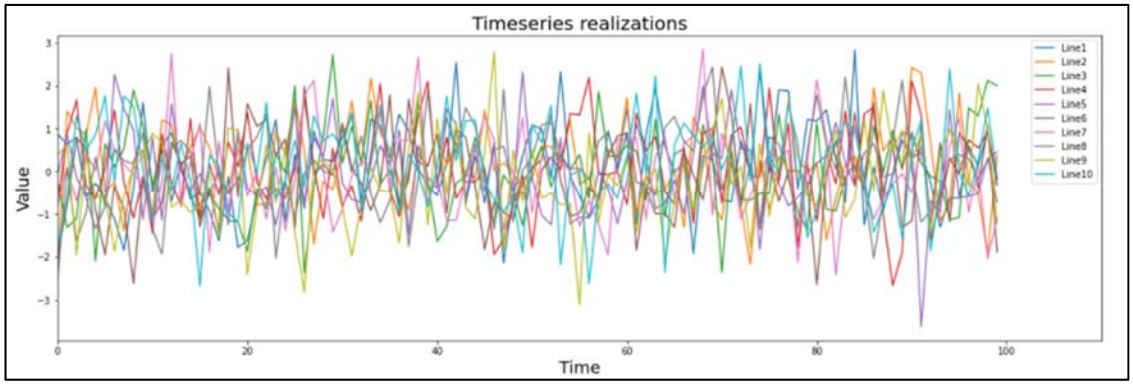

Figure 1. Time-series realized by sampling from the normal distribution.

Exercise 1.3

The autocorrelation and the auto-covariance are given from the following equations respectively:

$$R_x(t,\tau) = E[X(t) \cdot X(t+\tau)]$$

$$C_x(t,\tau) = R_x(t,\tau) - E[X(t)] \cdot E[X(t+\tau)] \rightarrow R_x(t,\tau) = C_x(t,\tau) + E[X(t)] \cdot E[X(t+\tau)]$$

Since the noise is uncorrelated, the auto-covariance:

$$C_x(t,\tau) = \begin{cases} \sigma_X^2, for\ all\ t, and\ \tau = 0 \\ 0, \quad for\ all\ t, and\ \tau \neq 0 \end{cases}$$

As a result, the autocorrelation formula can be updated to:

$$R_x(t,\tau) = \begin{cases} E[X(t)] \cdot E[X(t+\tau)], & for\ all\ t, and\ \tau = 0 \\ \sigma_X^2 + E[X(t)] \cdot E[X(t+\tau)], & for\ all\ t, and\ \tau \neq 0 \end{cases}$$

Exercise 1.4

1.4.a.

Since our realizations are Wide-Sense-Stationary (WSS), the autocorrelation is time invariant:

$$R_x(t,\tau) = R_x(\tau)$$

Hence, $R_x(0) = \sigma_X^2$, since the mean value of our realizations is zero.

The variance of the first realization is $R_x(0) = 1.1335$.

1.4.b

The variances were calculated for the remaining 9 realizations:

Realization	$\mathbf{Var[X_i] = R_{X_i}[0]}$	% difference with $\mathbf{R_{X_1}[0]}$
2	1.019	-2.2
3	0.729	-30.0
4	1.59	52.6
5	0.408	-60.8
6	1.312	25.9
7	1.173	12.6
8	1.211	16.2
9	0.953	-8.5
10	1.657	59.0

Table 1. Autocorrelation comparisons between the first realization and the remaining ones.

We notice that the variances of all realizations spread around the value of 1, which was the Variance of the initial Normal Distribution, from which all the realization points were derived. However, the differences of the autocorrelations have relatively large values for several realizations (e.g. realizations 8, 9).

In order to improve the estimation, we take advantage of the fact that we have multiple realizations of a WWS process. Combining these, we can predict the ergodic autocorrelation.

First, we need to prove that our process is ergodic, be showing that the ensemble mean for all the realizations is equal to the time mean. The above holds, since all the points were taken from the same distribution, with the same mean value.

Now, the autocorrelation can be calculated by the ergodic formula:

$$\hat{R}(k) = \frac{1}{N}\sum_{n=0}^{N-k-1} X[n]X[n+k],\ \ k = 0, \dots, N-1 \overset{k=0}{\Longrightarrow} \hat{R}(0) = \frac{1}{N}\sum_{n=0}^{N-1}(X[n])^2\ ,\ \ k = 0, \dots, N-1$$

Using the above formula, the ergodic autocorrelation is now $R_x(0) = 1.055$, while the autocorrelations for the remaining realizations and their % differences with the first realization are presented in the following table:

Realization	$\boldsymbol{Var[X_i] = R_{X_i}[0]}$	% difference with $\boldsymbol{R_{X_1}[0]}$
2	1.032	-8.4
3	1.056	-2.2
4	0.93	0.1
5	1.151	-11.8
6	1.084	9.1
7	1.012	2.7
8	1.183	-4.0
9	1.272	12.1
10	1.032	20.6

Table 2. Autocorrelation comparisons with ergodicity.

We immediately notice that the difference in the autocorrelation of the different realizations is decreased considerably, since it now has an absolute maximum value of 20.6%.

Exercise 1.5

1.5.a

The needed function was implemented in Python. The formula for the autocorrelation was used only for $k => 0$, while for $k < 0$, the graph was mirrored. The function calculates the autocorrelation for $-100 \leq k \leq 100$.

1.5.b

The plot for all 10 realizations can be seen below:

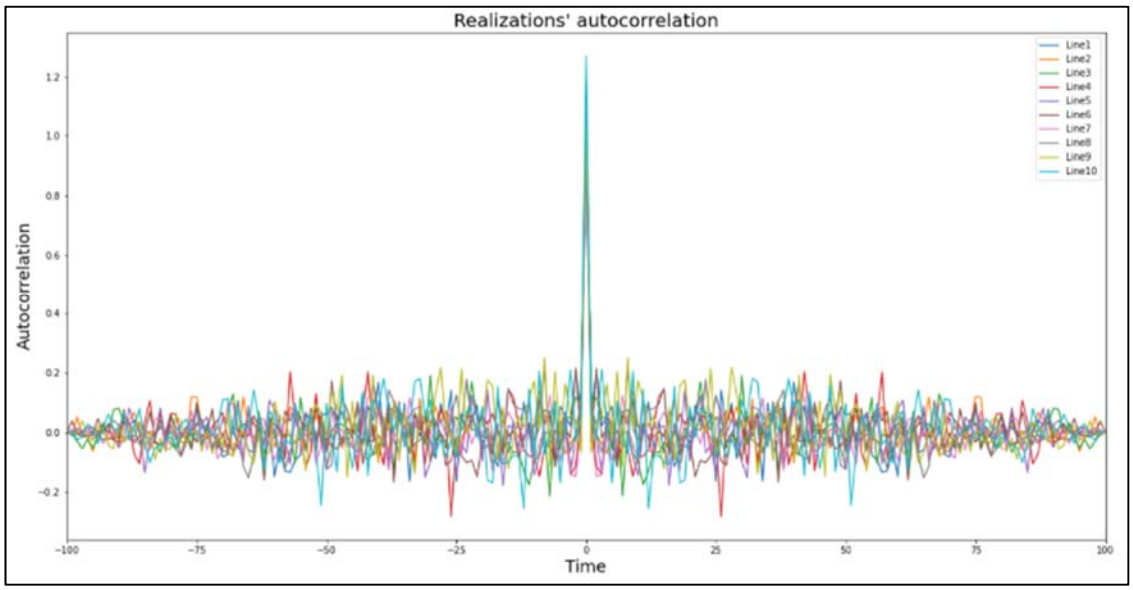

Figure 2. Autocorrelation graphs of the realizations. We notice that the autocorrelation decreases rapidly as we move away from zero.

1.5.c

If the autocorrelation function was a delta function, then $R_x(k = 0) = \delta(0) \rightarrow \infty$. But since $R_x(k = 0) = E[X^2(t)] = Average\ power\ of\ signal$, it is physically impossible that the average power in an infinite number. Also, the distribution is time discrete.

An example of a WSS signal with theoretically infinite average power is the Gaussian White Noise, which is used in signal filtering. However, after sampled in discrete time steps, the average power of that signal also gets finite.

1.5.d

Since $R_x(\tau)$ drops quickly right after $\tau = 0$, the samples have very little correlation with each other. In fact, judging by the way in which the data were constructed, it can be realized that there is no correlation between the data for values of $\tau \neq 0$, since no constraint was set for the structure of the each realization.

Hence, it can be concluded that except from the value $\tau = 0$, the estimation of autocorrelation can't be considered reliable, since the data were sampled in a completely random order.

Exercise 2

Exercise 2.1

Since we have many realizations of a WWS process, we can use the ergodicity autocorrelation formula that was given in exercise (1.5), we take the following plots for the 18 normal and the 10 ILD samples:

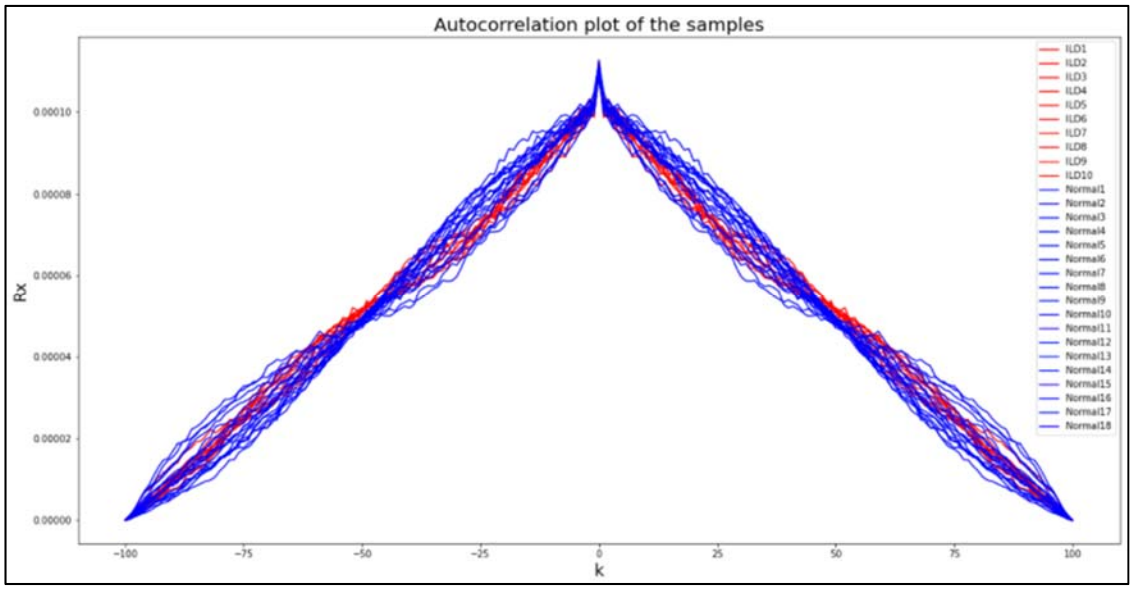

Figure 3. Autocorrelation estimates for the 28 labelled samples.

Exercise 2.2

In order to find the k*, in which the we can distinguish between normal and ILD samples, we look at the expected values of the correlations at each k. The optimal value k* will be the one where the absolute distance between the correlation expected value for Normal samples, and the correlation expected value for ILD samples is the maximum. In short:

$$k^* = argmax_k\left(\left|E\left[X_k^{normal}\right] - E\left[X_k^{ILD}\right]\right|\right)$$

We calculate that the value of k^* is 27, which can be visually justified by Figure 3.

Exercise 2.3

The plot containing the autocorrelation values for k^*can be seen below:

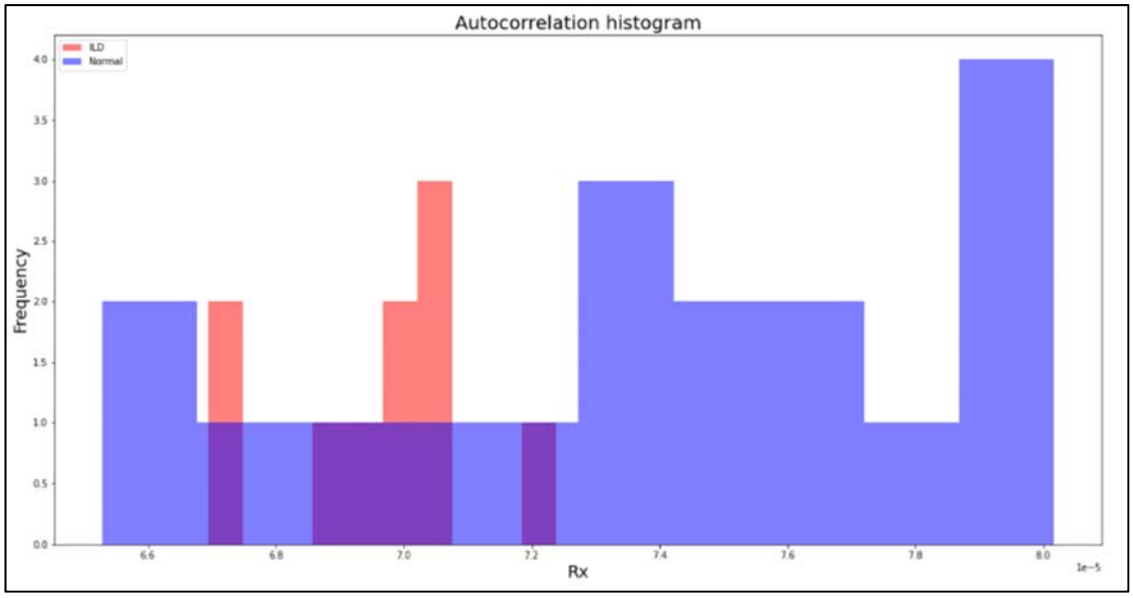

Figure 4. Autocorrelation histogram for k*=27.

It is clear that the values can't be perfectly classified.

Exercise 2.4

For the a priori probabilities, since the samples are representative of the real distribution, we calculate the probabilities using the following formulas:

$$P(L=0) = \frac{N_0}{N} = \frac{18}{28} = 0.643$$

$$P(L=1) = \frac{N_1}{N} = \frac{10}{28} = 0.357$$

,where N is the number of all samples, N_0 is the number of the samples labelled as "Normal", and N_1 is the number of samples labelled as "ILD".

Exercise 2.5

We calculate the expected value and variance for all the data points at k^*, for the "Normal" and the "ILD" dataset individually.

$$\mu_{k*}^{Normal} = \frac{1}{N_0}\sum_{i=1}^{N_0} x_{ik*}^{Normal} \qquad var_{k*}^{Normal} = \frac{1}{N_0}\sum_{i=1}^{N_0}\left(x_{ik*}^{Normal} - \mu_{k*}^{Normal}\right)^2$$

$$\mu_{k*}^{ILD} = \frac{1}{N_1}\sum_{i=1}^{N_1} x_{ik*}^{ILD} \qquad var_{k*}^{ILD} = \frac{1}{N_1}\sum_{i=1}^{N_1}\left(x_{ik*}^{ILD} - \mu_{k*}^{ILD}\right)^2$$

The final calculated results are:

$\mu_{k*}^{Normal} = 6.96\text{e}-05,\ var_{k*}^{Normal} = 2.33\text{e}-12$

$\mu_{k*}^{ILD} = 7.39\text{e}-05,\ var_{k*}^{ILD} = 2.11\text{e}-11$

The derived Gaussian distributions, together with the density histograms calculated in Exercise 2.3 can be seen below:

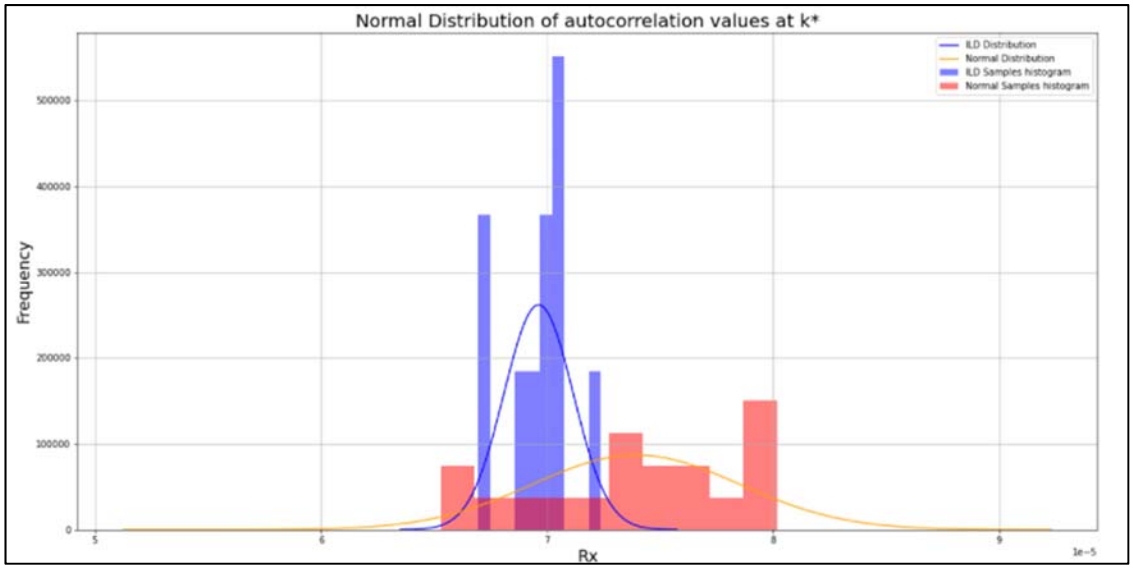

Figure 5. Normal Distribution of autocorrelation values at k*=27 compared to the respective histograms.

Exercise 2.6

In order to use the given MAP decision rule, we first need to calculate the probabilities $P(L = 0; R_X(k^*) = r)$ and $P(L = 1; R_X(k^*) = r)$. Using Bayes Rule, we have:

$$P(L = 0; R_X(k^*) = r) = \frac{P(R_X(k^*) = r; L = 0) * P(L=0)}{P(R_X(k^*) = r)}$$

$$P(L = 1; R_X(k^*) = r) = \frac{P(R_X(k^*) = r; L = 1) * P(L=1)}{P(R_X(k^*) = r)}$$

The probabilities $P(L = 0)$ and $P(L = 1)$ were calculated in exercise 2.4.

The probability $P(R_X(k^*) = r; L = 0)$ is given by integrating the Gaussian PDF with mean value μ_{k*}^{Normal} and variance var_{k*}^{Normal}. The probability $P(R_X(k^*) = r; L = 1)$ is given by integrating the Gaussian PDF with mean value μ_{k*}^{ILD} and variance var_{k*}^{ILD}, respectively.

Finally, the probability in the denominators are calculated as follows:

$$P(R_X(k^*) = r) = P(R_X(k^*) = r; L = 0) * P(L = 0) + P(R_X(k^*) = r; L = 1) * P(L = 1)$$

Taking all the above in account, we can now classify the samples from the "Normal" and "ILD" datasets:

Sample ID	Sample Classification	Result
1	Normal	True Negative
2	ILD	False Positive
3	Normal	True Negative
4	ILD	False Positive
5	Normal	True Negative

6	Normal	True Negative
7	Normal	True Negative
8	Normal	True Negative
9	ILD	False Positive
10	Normal	True Negative
11	Normal	True Negative
12	Normal	True Negative
13	ILD	False Positive
14	Normal	True Negative
15	ILD	False Positive
16	Normal	True Negative
17	ILD	False Positive
18	Normal	True Negative

Table 3. Normal Samples Classification.

Sample ID	Sample Classification	Result
1	ILD	True Positive
2	ILD	True Positive
3	ILD	True Positive
4	Normal	False Negative
5	ILD	True Positive
6	ILD	True Positive
7	ILD	True Positive
8	ILD	True Positive
9	Normal	False Negative
10	ILD	True Positive

Table 4. ILD Samples Classification.

The final classification Results can be summarized in the table below:

	Reality	
Prediction	**Normal**	**ILD**
Normal	12	2
ILD	6	8

Table 5. Estimations evaluation.

The correct estimates were 20, while the false ones were 8. The model has a 71.4% accuracy in the training set.

Exercise 2.7

Using the model derived from the previous exercises, we can now move on to the classification estimates of the signals of the file `test_scans.mat`.

We first calculate the autocorrelation function for the 10 samples:

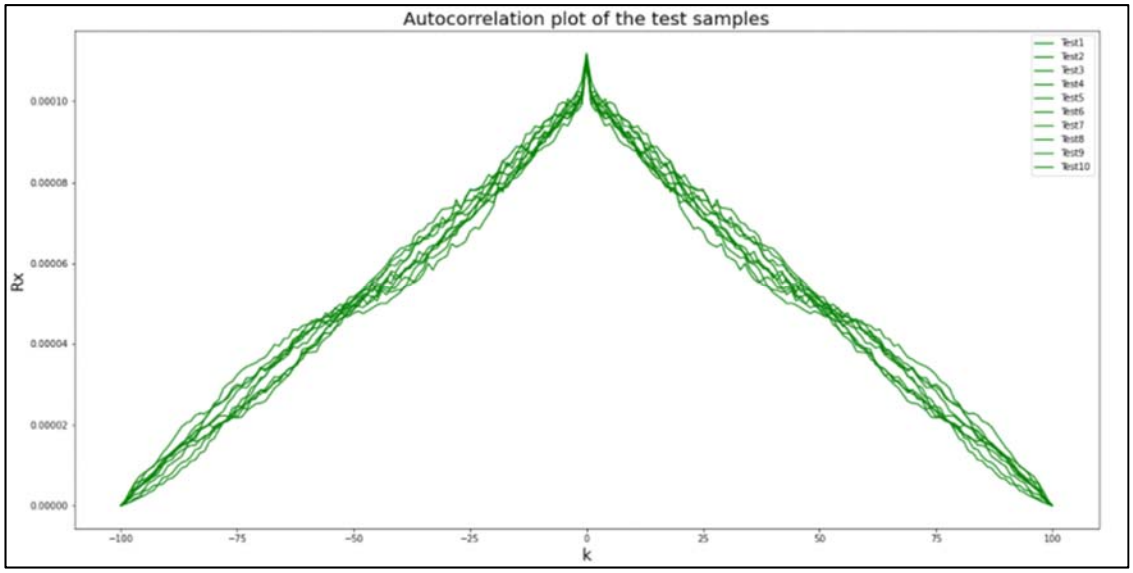

Figure 6. Autocorrelation estimates for the 10 test samples.

Finally, we can classify the samples by using the value $R_{X_i}(k^*)$, for each sample (i). The final classification results can be seen below:

Sample ID	Sample Classification
1	ILD
2	ILD
3	ILD
4	ILD
5	ILD
6	ILD
7	Normal
8	ILD
9	ILD
10	ILD

Table 6. Classification results for the test set.

A probably better solution

At this point, it is worth noting that the previous procedure was made for all possible values of k^*, in order to determine the optimal value, based on the classification accuracy in the training set. After the iteration, it was realized that using $k^* = 22$, the final accuracy is 78.6%.

Hence, for completeness, the results for Exercises 2.3, 2.5, 2.6 & 2.7 are calculated again, using $k^* = 22$.

- The autocorrelation histograms for "Normal" and "ILD" sets:

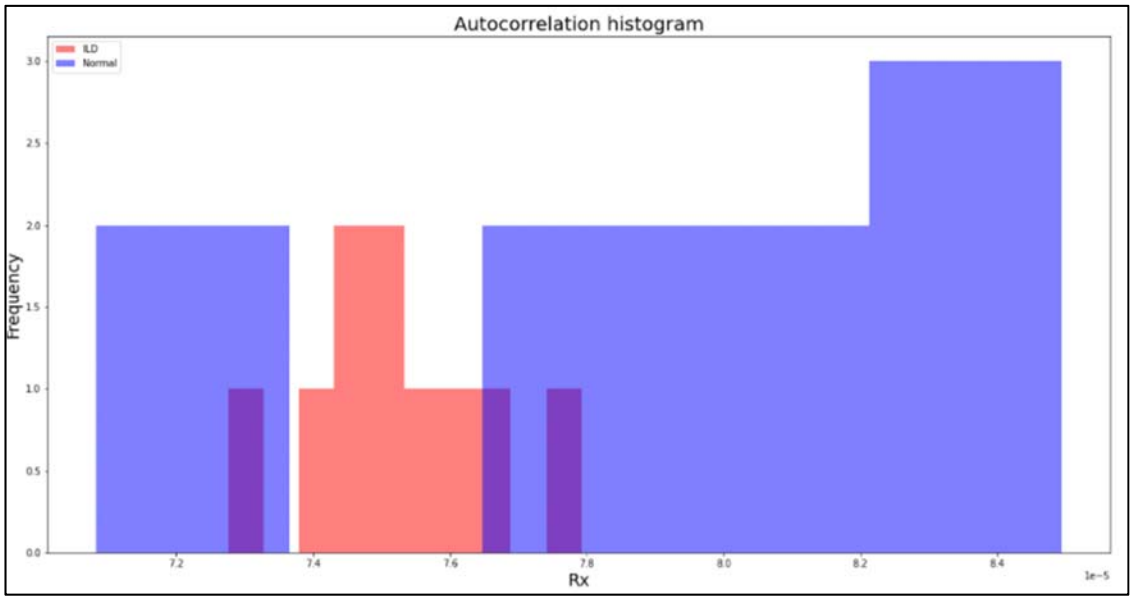

Figure 7. Autocorrelation histogram for k*=22.

We can see that the two classes can be distinguished much easier.

- The Normal Distributions of the autocorrelations for the two sets:

The derived expected values and variances:

$$\mu_{k*}^{Normal} = 7.52\text{e} - 05, \qquad var_{k*}^{Normal} = 1.93\text{e} - 12$$

$$\mu_{k*}^{ILD} = 7.92\text{e} - 05, \qquad var_{k*}^{ILD} = 2.02\text{e} - 11$$

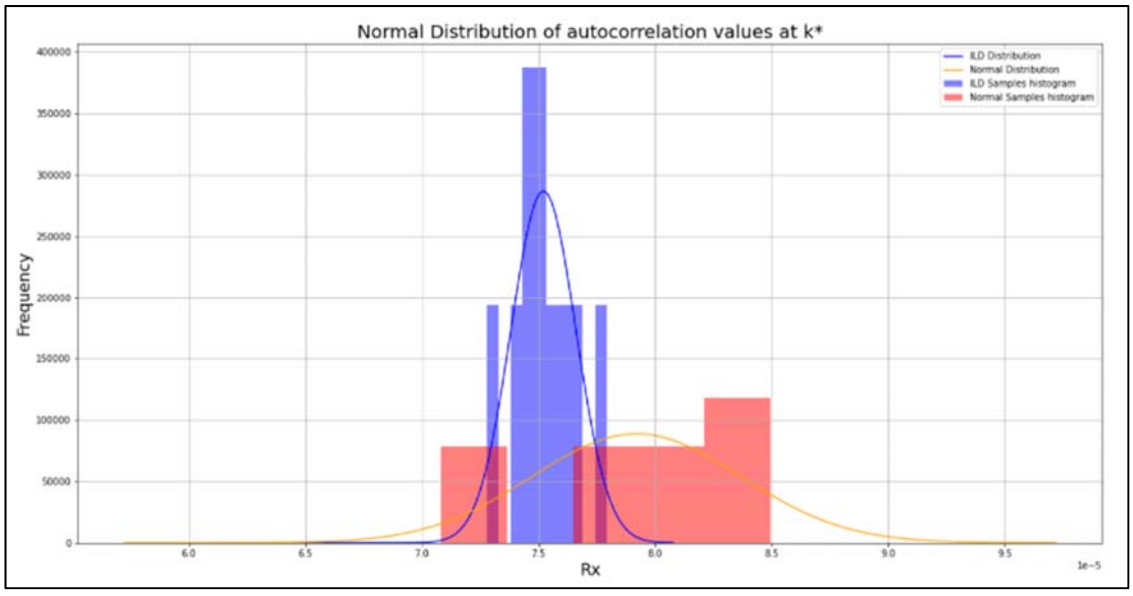

Figure 8. Normal Distribution of autocorrelation values at k*=22 compared to the respective histograms.

- The classification evaluation of the training set:

Sample ID	Sample Classification	Result
1	Normal	True Negative
2	ILD	False Positive
3	Normal	True Negative
4	ILD	False Positive
5	Normal	True Negative
6	Normal	True Negative
7	Normal	True Negative
8	Normal	True Negative
9	Normal	True Negative
10	Normal	True Negative
11	Normal	True Negative
12	Normal	True Negative
13	Normal	True Negative
14	Normal	True Negative
15	ILD	False Positive
16	Normal	True Negative
17	ILD	False Positive
18	Normal	True Negative

Table 7. Normal Samples Classification for k*=22.

Sample ID	Sample Classification	Result
1	ILD	True Positive
2	ILD	True Positive
3	ILD	True Positive
4	Normal	False Negative
5	ILD	True Positive
6	ILD	True Positive
7	ILD	True Positive
8	Normal	False Negative
9	ILD	True Positive
10	ILD	True Positive

Table 8. Normal Samples Classification for k*=22.

	Reality	
Prediction	**Normal**	**ILD**
Normal	14	2
ILD	4	8

Table 9. Estimations evaluation.

- Predictions on the test set:

Sample ID	Sample Classification
1	ILD
2	ILD
3	ILD
4	Normal
5	ILD
6	ILD
7	Normal

8	ILD
9	ILD
10	ILD

Table 10. Classification result of the test set for k*=22