Abstract — An early diagnosis of human cancer is of crucial importance for successful therapies. Therefore cancer diagnosis via ESR (Electron-Spin-Resonance) spectroscopy provides a new promising approach. A 33-dimensional vector describes a state, i.e. the functional properties of albumin in the blood. This vector is derived from the ESR-Spectrum of a human serum sample mixed with so-called spin-probes. These spin-probes are bound by albumin. In case of cancer some functions of albumin cause changes in the ESR-Spectrum and thus the vector, which enables differentiation between the classes "healthy" and "suspicious" [1], [2]. It is assumed that this vector complies with a Gaussian distribution. The distribution of "healthy" and "suspicious" patients can be splitted into multiple distributions using the EM (Expectation-Maximization [3]) method. This leads to a more accurate description of the system as well as a decrease of the classification error rate. The focus is on the overlapping area of the "healthy" and "suspicious" distributions, where the classification error rate has its maximum. This EM method is effected in three different modes: splitting with full covariance, diagonal or identity matrix. The variations are analyzed with respect to their classification performance and in combination with LDA (Linear-Discriminant-Analysis [4]) to reduce matrix dimensions. The classification compares the maximum values achieved for each Gaussian distribution ("healthy" and "suspicious") where the class with the higher value is chosen. In addition the analysis is repeated after transformation of normal distributed data to standard normal distributed data. The respective results were compared. Using these methods in combination should decrease the error rate of classification of the two-class-problem ("healthy" and "suspicious").

Index Terms — ESR Spectroscopy, MMS-Test, Early Cancer Recognition, Expectation-Maximization-Estimation, Linear-Discriminant-Analysis.

I. INTRODUCTION

CANCER is one of the most common diseases. The best anticancer strategy is early detection of malign processes in the human organism. The MMS-Test (mobility of molecular structures) offers the opportunity to detect cancer independent from its type and location. Therefore, the albumin transport parameters are determined using electron spin resonance spectroscopy. One major task of serum albumin is the transport of hydrophobic substances, like fatty acids and vitamins as well as toxins and several drugs. During transport of hydrophobic substances serum albumin changes its molecular conformation in a specific manner [5]. In addition, albumin plays an important role in modulating the serum concentration of different low molecular weight ligands and phospholipids produced by tumor cells [6]. These cancer-derived proteins are present in the blood circuit in very low concentrations. Albumin as the major transport protein in the human organism protects these cancer markers from catabolism and significantly amplifies their concentration. It is scientifically confirmed that these proteins block the albumin function by changing its conformation. The conformational change of serum albumin, and thus its transport and detoxification properties, can be determined with the help of the MMS-Test [7].

In this study the results of the MMS-Test are analyzed with Expectation-Maximization-Estimation of Mixture-Densities. Therefore the MMS-Test results of 1176 patients were analyzed.

II. METHODS

A. MMS-Test

The ESR spectroscopy signal is excited by molecules containing unpaired electrons. These compounds are hazardous to health. However, the human immune system is able to reduce the concentration of these molecules. The need of unpaired electrons for ESR measurements can be compensated using spin-probes. On the basis of ESR spectroscopy a special method was developed for the determination of the transport and detoxification properties of the serum albumin, the MMS-Test. Conformational changes are induced in vitro simulating different transport situations by addition of different concentrations of a certain polar reagent. The ability of albumin to realize these conformational changes is characterize by means of a particular label. This label (16-doxyl stearic acid) is specific for the protein. The affinity of albumin for 16-doxyl stearic acid is exactly identical its affinity for unlabeled stearic acid [7]. The resulting ESR-Spectrum is analyzed using mathematical simulation methods. The simulation is performed using least-square fitting between a model spectrum and the spectrum that is measured experimentally. The ESR-Spectra generated by stearic acid spin-probes bound to albumin, reveals four distinct spectral components. The two major components (figure 1, lines B and C) are derived from the fatty acid spin-label bound to albumin. The two remaining components are derived from free or unbound
fatty acids in the solution (line D). The latter can also occur organized into a cluster of fatty acid micelles (line E). The values of ideal spectrum parameters if defined by mathematical simulation. These parameters represent the equation that provides the best curve fit of the simulated and measured spectra. They include the intensity of each spectral component as well as the specific ESR parameters, which determine position, width and shape of spectrum lines. Each ESR-Spectrum shows the structural and functional characteristics of the protein. The binding of spin-probes to albumin reflects these characteristics. Therefore, three 50 µl aliquotes of serum or plasma samples are mixed with different concentrations of 16-doxyl stearic acid and ethanol. This mixture is incubated with continuous shaking for 10 min at 37°C. Thereafter the mixtures are filled into capillaries and transferred to the ESR analyzer, where the spectra are recorded automatically. These spectra are analyzed by a purpose-implemented software.

Fig. 1. Typical ESR-Spectrum of a spin-probe bound albumin (line A) and its components (lines B - E). (A) Composite ESR-Spectrum of 16-doxyl stearic acid (16-DS) bound to albumin. (B) Spectrum corresponding to 16-DS bound to primary binding sites on albumin. (C) Spectrum corresponding to 16-DS bound to secondary binding sites on albumin. (D) Spectrum corresponding to unbound 16-DS. (E) Spectrum corresponding to micelles of 16-DS.

**B. Data and Algorithm**

The database contains 1176 patients in terms of 33-dimensional vectors. These vectors are the results obtained by the analysis of the ESR-Spectra of the serum of the patients. The database contains datasets of patients with cancer and chronic disease as well as of healthy people. (see table I)

<table>
<thead>
<tr>
<th>status</th>
<th>absolute number</th>
<th>relative number</th>
</tr>
</thead>
<tbody>
<tr>
<td>healthy</td>
<td>550</td>
<td>46.77%</td>
</tr>
<tr>
<td>cancer</td>
<td>498</td>
<td>42.35%</td>
</tr>
<tr>
<td>chronic disease</td>
<td>128</td>
<td>10.88%</td>
</tr>
<tr>
<td>∑</td>
<td>1176</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

**TABLE I**

DATABASE STRUCTURE

The database is divided into three almost equal parts for classification.

- 1/3 training data
- 1/3 test data
- 1/3 validation data

Every third dataset of the database is assigned to the validation data. Whereas the training and test data are newly-arranged for every run of the program.

The Linear-Discriminant-Analysis (LDA) can be used for reduction of matrix dimensions. Therefore, a coordinate transformation is realized which leads to a better separation of the classes [4].

It is possible to accomplish two different pre-processing steps. Whereas, either one or both of these methods can be applied. Firstly, the pre-processing can be achieved by normalization of the data and secondly, by LDA to reduce matrix dimensions. The LDA is applied to the training data and assigned to the test data. After the pre-processing steps the set of the training data is divided in "healthy" and "suspicious" data.

The Expectation-Maximization (EM) procedure can be used to split high dimensional Mixture-Densities. During EM method, the split of "healthy" and "suspicious" takes place separately. Before starting the EM method, an initial split has to be performed. That occurs by means of the "k.means" method. Therefore two randomly chosen vectors of the distribution are selected as the new center of the splitted distribution. All further vectors are shifted to one of the initial vectors in accordance to their respective distance. As a result, two new distributions are received, which will then be estimated again. After that every further vector will be newly evaluated and shifted into the respective distribution. This classification is carried out according to the following equation for multivariate normal distribution:

\[
 f(\bar{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\Sigma}} e^{-\frac{1}{2}(\bar{x}-\mu)^t \Sigma^{-1}(\bar{x}-\mu)}
\]

These two steps will be repeated iteratively until no further change occurs within the distribution. A reassignment of data to the distributions is a simplification of the EM method. This method is called Viterbi-Approximation. In this case it is assumed that the weighting factor (\(\alpha\)) is constant.

\[
 p(x|G) \approx \max_{i=1}^{n} \alpha_i p(x|G_i)
\]  

(1)

During the process of classification every vector of the test data is assigned to the highest score calculated by the function for multivariate probabilities. Hence, assigned either to the "healthy" or the "suspicious" group. After classification of the whole dataset the error rate is calculated. This is repeated several times following the calculation of the averaged error rate. Tests show that 100 repetitions are enough to get reliable results.

In the course of this study, several analysis with different parameters were conducted. More precisely, the type of the matrix was varied as well as the subsequent settings of the LDA. The best result of every analysis was validated using the validation data.
III. Experiments

A. Expectation-Maximization

This study deals with the analysis of the two class problem. The classification is effected into the "healthy" and the "suspicious" group. The "suspicious" group consists of the patients with cancer and chronic disease. Table II shows the results after classification. The best results could be obtained without EM split using the full covariance matrix.

<table>
<thead>
<tr>
<th></th>
<th>no split error rate in %</th>
<th>one split error rate in %</th>
<th>two splits error rate in %</th>
</tr>
</thead>
<tbody>
<tr>
<td>VK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>norm.</td>
<td>14.54</td>
<td>14.97</td>
<td></td>
</tr>
<tr>
<td>VK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>norm.</td>
<td>14.63</td>
<td>15.08</td>
<td>15.15</td>
</tr>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>norm.</td>
<td>17.88</td>
<td>15.54</td>
<td></td>
</tr>
<tr>
<td>DM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>norm.</td>
<td>17.91</td>
<td>15.63</td>
<td>15.66</td>
</tr>
<tr>
<td>IM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>norm.</td>
<td>17.72</td>
<td>15.88</td>
<td></td>
</tr>
<tr>
<td>IM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>norm.</td>
<td>17.88</td>
<td>15.88</td>
<td></td>
</tr>
</tbody>
</table>

TABLE II
ERROR RATE OF CLASSIFICATION OF TEST DATA WITH FULL COVARIANCE MATRIX (VK), DIAGONAL MATRIX (DM) AND IDENTITY MATRIX (IM)

H - healthy, C - suspicious and norm - normalized

As appears from table II, there is a significant difference of the error rate of the "healthy" and the "suspicious" group using the diagonal or identity matrix. As the classifications is carried out according to the following formula:

\[
\frac{f_{\text{healthy}}(x)}{f_{\text{suspicious}}(x)} > \alpha \text{ then healthy, else suspicious}
\]

it is possible to modulate the error rate difference between the two groups changing the operation point \(\alpha\). In a normal classification the operation point equals one. Hence, one goal was the optimization of this operation point.

It appears from figure 2 that using the identity matrix the optimal operation point regarding the total error rate is \(\alpha = 1.00\). Figure 2 shows the resulting curve of the total error rate over the shift of the operating point \(\alpha\).

The best operation point for the diagonal matrix is \(\alpha = 10.24\) (data not shown).

B. Expectation-Maximization & LDA

Another part of this work is the introduction of the LDA to classification. Here the matrix dimension is reduced stepwise from 33 down to one dimension. Within each of these steps up to three EM splits are carried out. Figure 3 shows the total error rate after classification related to the matrix dimension using the identity matrix.

As appears from the figure the lowest error rate after the classification is 12.94%. This error rate is obtained at two dimensions and with one EM split.

With the full covariance matrix the lowest error rate after classification using LDA was 13.51% (data not shown).

IV. Discussion

A. Expectation-Maximization

The results show, that in this case the lowest error rate after classification could be achieved with the full covariance matrix (table II). The higher error rates from the diagonal and identity matrix indicate that these types of matrices are not suitable for classification in the high dimensional space. After splitting, the results obtained using diagonal and identity matrix could be improved. However, the respective error rates are still higher than the one obtained using the full covariance matrix. Furthermore, for the DM and the IM, the results show a significant lower error rate for the "healthy" group. A combination of diagonal matrix for the "healthy" group and full covariance matrix for the "suspicious" group leads to an error rate of 28.47%, which is an explicit degradation of the previous results. Besides, the normalization of the datasets could not improve the results significantly. Shifting the operation point \(\alpha\) improved the results of the diagonal matrix but with an error rate of 16.87% it is still higher than the error rate of the full covariance matrix.

B. Expectation-Maximization & LDA

The reduction of matrix dimension lead to a decrease of the error rate after classification. The major Eigen value is \(10^{12}\) times higher than all other Eigen values. This means that 99.99% of information are contained in this dimension.
It appears from figure 3 that the error rates after LDA are higher than without LDA at 33 dimensions. The error rates increase after LDA if identity matrix is used. This is due to the fact that the weighting factor of the dimensions is not taken into account. Therefore, the identity matrix is not suitable for high dimensional distributions. After removing the datasets of chronic diseased patients from the "suspicious" group the total error rate decreased to 8.51% after LDA (one dimension) and EM split (one split). A possible reason for this change, is that chronic diseases influence the serum albumin in different way than cancer does.

V. Conclusion

The best classification result is obtained using LDA with 2 dimensions and the identity matrix with a total error rate of 12.94% on the test data and 12.93% on the validation data. The EM method improved the results slightly. In combination with the LDA the EM method gives better results but only up to two splits. Thereafter the total error rates increase. As the calculation is rather complex, the EM method is not working at the full covariance matrices. Table III shows the best total error rates of test and validation data for different methods.

<table>
<thead>
<tr>
<th>method</th>
<th>error rate (test data)</th>
<th>error rate (validation data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>EM[0]</td>
<td>14.54%</td>
<td>13.37%</td>
</tr>
<tr>
<td>LDA[1] + EM[0] (VK)</td>
<td>13.51%</td>
<td>12.99%</td>
</tr>
<tr>
<td>LDA[1] + EM[1] (DM)</td>
<td>13.68%</td>
<td>12.43%</td>
</tr>
</tbody>
</table>

TABLE III
ERROR RATE OF TEST AND VALIDATION DATA FOR DIFFERENT METHODS

VI. Outlook

The lower error rates with diagonal and identity matrix indicate that these vectors are not normally distributed. The Kolmogrov-Smirnov-Test could for instance be used to analyze the distribution of the vectors. Another approach is to perform a Principal Component Analysis (PCA) to reduce the matrix dimension.

References