Acknowledgements

I wish to express my sincere appreciation to Professor Michalis Zervakis for his valuable assistance, his patience, the excellent cooperation in every single stage of this work, and his friendly support throughout all the years of my studies.

I would also like to thank my friends and PhD students Ioanni Dimou, Mixali Kounelaki, and Mixali Blazantonaki for the excellent cooperation we had all these years. In addition, special thanks to Professor Antonina Starita (Dip. di Informatica, Pisa), Professor Elia Biganzoli (Cancer Institute of Milan), Flavio Baronti (Dip. di Informatica, Pisa), and Alessandro Passaro (Dip. di Informatica, Pisa) for their valuable support during my stay in Italy.

Finally, I would like to thank Associate Professor Athanasios Liavas and to Associate Professor Evripides Petrakis for their participation in the supervisory committee.

This thesis is dedicated to my lovely family and to Iliana. I am truly grateful to them for their love, patience, and understanding.

Present work was supported in part by Biopattern, IST EU funded project.
www.biopattern.org

Contents

Figure Contents

Table Contents

1 Introduction

Acute myeloid leukemia (AML), also known as acute myelogenous leukemia, is a cancer of the myeloid line of white blood cells, characterized by the rapid proliferation of abnormal cells which accumulate in the bone marrow and interfere with the production of normal blood cells. AML is the most common acute leukemia affecting adults, and its incidence increases with age. Although AML is a relatively rare disease, accounting for approximately 1.2% of cancer deaths in the United States, its incidence is expected to increase as the population ages.

The symptoms of AML are caused by replacement of normal bone marrow with leukemic cells, resulting in a drop in red blood cells, platelets, and normal white blood cells. These symptoms include fatigue, shortness of breath, easy bruising and bleeding, and increased risk of infection. Although several risk factors for AML have been identified, the specific cause of AML remains unclear. As an acute leukemia, AML progresses rapidly and is typically fatal within weeks or months if left untreated.

Acute myeloid leukemia is a potentially curable disease; but only a minority of patients is cured with current therapy. AML is treated initially with chemotherapy aimed at inducing a remission; some patients may go on to receive a hematopoietic stem cell transplant. Areas of active research in acute myeloid leukemia include further elucidation of the cause of AML. Many cancer research centers and scientists focused on the identification of better prognostic indicators, development of new methods of detecting residual disease after treatment, and the development of new drugs and targeted therapies.

Recent studies have deployed a thorough examination in which a specific type of indicators, named as cytogenetic abnormalities, is considered as the most valuable prognostic determinants in acute myeloid leukemia (AML) **[1]**, **[2]**, **[3]**. Over the last years, DNA microarray experiments are being used to gather information from tissue and cell samples regarding gene expression differences that will be useful in diagnosing disease. The gene expression profile is significant information for the determination of the AML prognosis and many researches have taken into consideration the importance of this profile to the prognosis of the AML disease **[4]**, **[5]**, **[6]**, **[7]**.

The aim of the current study was the evaluation and benchmarking of various data analysis methods, through the field of pattern recognition, and the prospect of exploiting

knowledge therein. Supervised learning, a procedure in which individual items are placed into groups based on quantitative information on one or more characteristics inherent in the items, was adopted in our study. The application of the classification technique to the AML, yields promising results for an accurate prognosis of the disease and a reliable and strict evaluation of the indicators participating to the classification.

The dataset relative to the AML99 protocol was provided by the Gimema group **[8]**, an Italian group which has been expertly involved with the research of the Adult Hematological Diseases with coordination among more than 100 centers for uniform treatment protocols and data collection.

The Comparison of different classification approaches to the AML dataset, based on their potential and limitations, was pursued and analyzed. Modern classification algorithms based on the Support Vector Machines (SVMs), the Least-Squares Support Vector Machines (LS-SVMs) and many others classifiers were applied to the biomedical data, with the purpose of predicting the outcome probabilities over follow-up.

Precisely, the overall study was separated into several parts. When a dataset was entered the classification procedure, several classifiers like the Support Vector Machines and a number of modifications and extensions related to the SVM methodology were applied to the dataset. This process, named by our group as Level 1 classification, was the level in which all classifiers participated individually to the problem, attempting to achieve the best classification result. Widely used classifiers like the Probabilistic Neural Networks, Fisher Discriminant Function, Radial Basis Neural Network Classifier and many others contributed to this level with their prediction accuracy. The evaluation of the Level 1 classifiers was assessed by several measures like the Region of Convergence (ROC), the Specificity and many others that will be analytically discussed below.

Our research was also focused on the examination of the most relevant indicators from the tested dataset that contributed to the improvement of the prognosis. For that reason, the implementation of feature selection techniques was deemed essential. Feature reduction methods, based on wrapper and filter techniques, aimed to improve the performance of the classification by selecting the most highly ranked indicators that provided the best classification accuracy.

We will analytically explain below that this research had to deal with several datasets, coming from the raw data, instead of having a unique set for examination. All the available datasets were classified at Level 1, and the one that provided the best classification accuracy was further examined according to the feature selection methodology. This implementation focused on the achievement of a better or at least equivalent classification performance by having fewer indicators in the dataset.

Afterwards, the classification results from the individually operated classifiers functioned as a meta-dataset for classification. Level 2 classification, widely named as fusion of the classifiers or classifiers ensemble, was the phase in which a group of classifiers cooperated in order to achieve an improved classification performance. Several pairwise and non-pairwise measures functioned significantly during Level 2 with the purpose of verifying the diversity degree of all the possible groups of classifiers. The criterion for choosing a group of classifiers for the fusion procedure will be discussed in the following chapters. Modern classifier ensemble methods like the Decision Templates and Naïve Bayes techniques were implemented and analyzed.

At the end of Level 1 and Level 2 classification, a survival analysis was set up for benchmarking the supervised learning results from the statistical point of view. Survival analysis is a considerable criterion for prediction and evaluation analysis, an analysis which is widely used by the clinicians. For that reason, the application of this analysis was essential in our study.

Before we proceed with the presentation of our research, it is important to do a brief report with the problems that we had to overcome, separated into theoretical and technical problems. From the theoretical point of view, the current research was a very difficult and not a familiar subject for us. This study started from scratch, and the lack of a good theoretical background on biological and hematological problems was an issue for our group in order to understand the meaning and the significance of each indicator. The Gimema's trials produced a set of many indicators separated into different biological types, from clinical analysis results to gene's expression profiles. What is more, the entire treatment was composed of many and complicated phases, starting from the initial examination of a patient to the bone marrow, when needed, transplantation.

Our team had to decode the AML99 protocol, to specify the phases required for our analysis, and to choose the most significant indicators for the supervised learning procedure. Hopefully, we had a relevant support from the doctors participating to this project and their guidelines helped us during the phases this study. In the following chapters, we are going to give an analytical presentation of the problem and the solutions that we chose.

Technically, the entire procedure was time-consuming with a really high computationally cost. Despite the fact that much attention was paid for the generation of fast and easy structured algorithms, the need for an exhaustive evaluation of the classification techniques caused this event. All the simulations were implemented in our laboratories under the supervision of professor Michalis Zervakis.

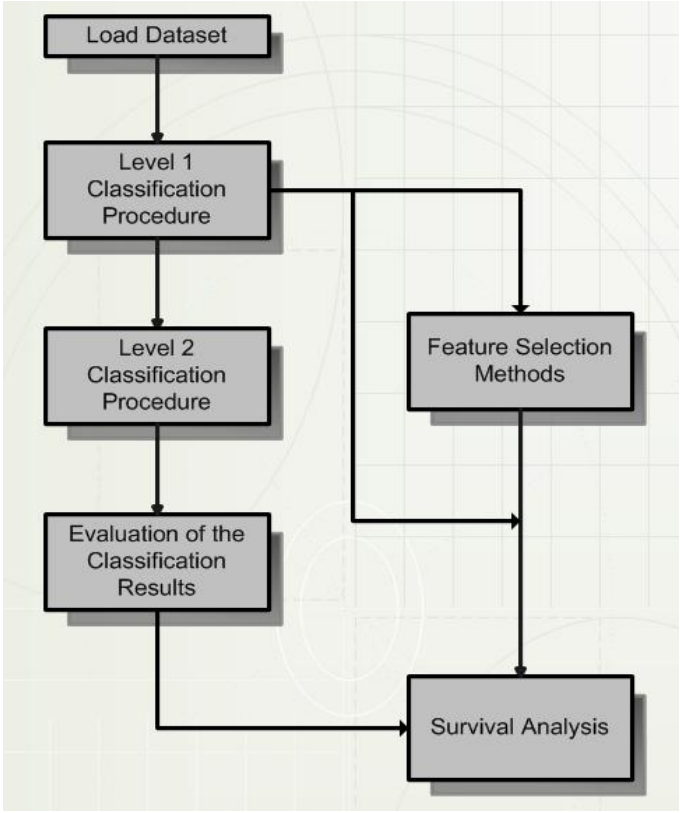

Figure 1 Graphical representation of the research

2 Preprocessing

2.1 Understanding the dataset

Our first aim, before starting with the implementation of the supervised learning techniques, was to derive the information obtained from the Gimema protocol and assess the significance of its clinical and biological indicators. For that reason, thanks to the Biopattern group for its support, our group urgently visited the University of Pisa and the Cancer Institute of Milan. Our group, together with the clinicians from Rome and our coordinators from Pisa and Milan, finally succeeded to understand the possible issues arising from the AML99 protocol, to decipher the information provided by the dataset, and to define the oncoming working fields which are depicted below.

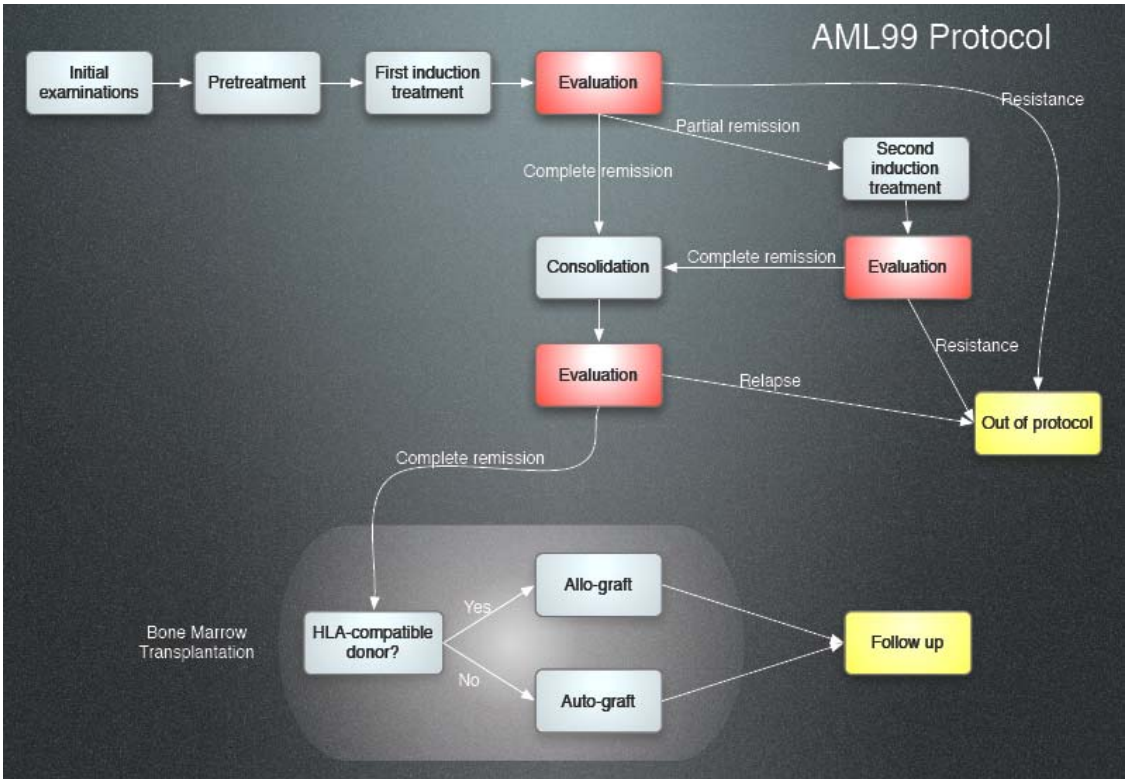

Figure 2 Schematic illustration of the AML99 protocol

According to Figure 2, before starting the Induction Course (chemotherapy) patients were supposed to receive a pre-treatment therapy, intended to decrease White

Blast Cells (WBC) and keep the disease under control. Meanwhile, several cytogenetics and molecular biology markers were measured and analyzed.

Referring to the AML99 Protocol, the evaluation of the response to the induction treatment for Acute Myeloid Leukemia (AML) was scheduled between 31 and 38 days from the beginning of the induction treatment. At this stage patients were categorized as those who achieved Complete Remission of the disease (CR), patients achieved Partial Remission (PR), patients still affected by the disease with no improvement at all (Resistant), and those who didn't succeed to defeat the disease (death in induction). Afterwards, patients in PR entered a second induction treatment. The definitive evaluation of the response to the induction treatment should happen at about 80 days. This term should (could) be procrastinated at 90 days for accounting of a further shift induced by a possible cardiac or hematological toxicity. The only type of failure which could be recorded at its actual time was the Induction Death (ID), whereas the achievement of Complete Remission (CR) or of Resistance was subordinated to the time of the evaluation.

The outcome for adults with AML depended on a variety of factors including age of the patient, history of the patient, intensity of post-remission therapy, and biologic characteristics of the disease, the most important of which was the information from the possible cytogenetic abnormalities of a patient at presentation.

As described above, patients affected by the disease followed a two-cycle treatment. After the first induction cycle patients were analytically categorized, regarding their response to the treatment, into several categories outlined in Table 1. Patients in Partial Remission (PR) after the first course were supposed to have a second induction cycle. After the second induction cycle, a patient was evaluated either as responder or non-responder (or dead etc, but no PR anymore). The non-responders who didn't correspond to the treatment were off-protocol. Then, responders had a consolidation cycle and an allogenic or an autologous transplant, depending on donor availability (see Figure 2 for further details). The overall response for a patient (at second cycle if done, otherwise at first cycle) is represented below.

response to first induction treatment		
Outcome Response	**# of Patients**	**Percentage**
Complete Remission	**347**	**68.17**
Partial Remission	**4**	**0.79**
Resistant	**88**	**17.29**
Extra-medullar loc.	**2**	**0.39**
Early death	**11**	**2.16**
Hypoplastic death	**56**	**11.00**
Hypoplasia	**1**	**0.20**

Table 1 Response after first induction cycle of the protocol.

The complexity and the correlation, in some cases, among the possible outcome responses (for example, strong relation between Early Death and Hypoplastic Death) were problems that could be easily affect the classification performance. The clinicians gave the solution to this problem, by suggesting the reduction of the outcome responses into only three possible outcomes. Outcome response "**Resistance"** consisted of cases "**Resistant**", "**Partial Remission**" and "**Extra-medullar loc.**", whereas response "**Induction death**" was composed of cases "**Early death**" and "**Hypoplastic death**". The new classes are outlined in Table 2.

response after the overall induction treatment		
Overall Outcome Response	**# of Patients**	**Percentage**
Complete Remission (CR)	**347**	**68.17**
Resistance	**95**	**18.66**
Induction death	**67**	**13.16**

Table 2 Outcome responses after the overall induction treatment

A survival analysis was set up on a discrete time basis by our coordinators in Milan **[9]**, following a partition of the time axis, using as a rationale for the discretisation

the time points indicated by the protocol for the clinical evaluation of the response ([0, 30), [30, 60), [60, 90), [90,...)). Afterwards, the survival function within each interval was examined (see Figure 3). The last proposal made possible to lead a useful comparison between a suitable survival regression model for discrete time intervals and a possible (supervised pattern recognition) solution to a probabilistic classification problem.

The curves represented the probability of the response as the first occurring event in presence of the possibility of the occurrence of the other events. The largest numbers of non-ID events were recorded between 30 and 60 days from diagnosis (at the end of first induction treatment) and successively between 60 and 90 days.

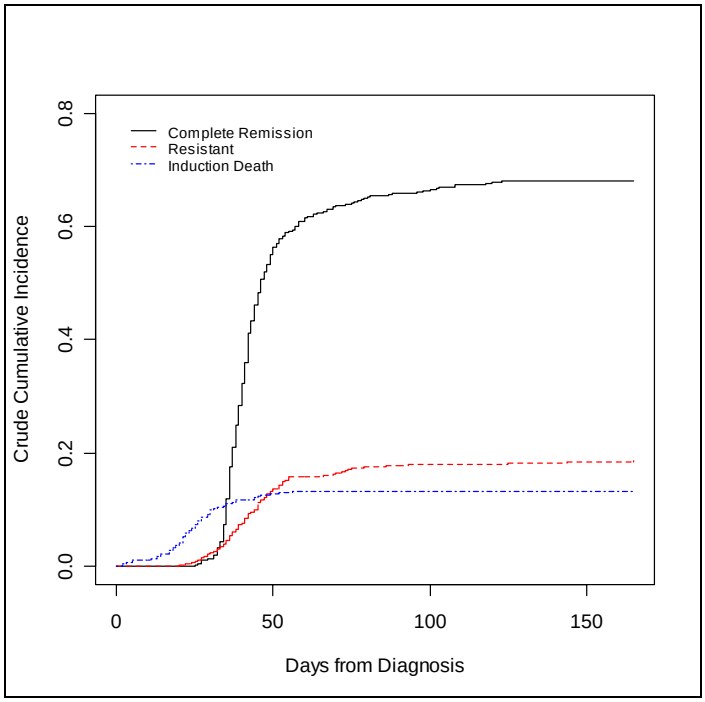

Figure 3 Crude Cumulative Incidence [9]

The survival analysis pointed out the classification study into definite time periods of diagnosis. Particularly, our research focused, respectively, on the prediction of the events after the Short Term and Long Term Analysis. In discrete time interval [0, 30), and specifically at time 30, we observed a significant difference between the cumulative incidence of response "**Induction Death**" and responses "**Complete Remission**" and "**Resistant**". That time period was named as Short Period of analysis. This difference

motivated us to study that time period and to perform a binary classification to classes "**Induction Death**" and "**all others**", where class **"all others"** was the union of outcomes "**Complete Remission**" and "**Resistant**".

On the other hand, Long Term Analysis was the analysis carried out after the completion of the two-cycle, when needed, induction treatment (after 60 days of treatment). A major difference between survival curve "**Complete Remission**" and the curves from the responses "**Induction Death**" and "**Resistant**" was noticeable at time $t = 60$. For that reason, and based on the assumption that this event might be a promising criterion so far for a good prognostic system, a binary classification problem between classes "**Complete Remission**" and "**all others**" was held. Class "**all classes**" composed of responses "**Induction Death**" and "**Resistant**".

When the Long and Short Term Analysis were under examination, the classification problem focused on the achievement of Complete Remission and the accurate prediction of event "**Induction Death**", respectively. The following tables depict the two binary classification problems.

Complete Remission (CR)			
CR	**Details**	**# of Patients**	**Percentage**
1	**CR achieved**	**347**	**68.17**
0	**No CR achieved**	**162**	**31.83**

Table 3 Classes at Long Term Analysis

Induction Death			
inde	**Details**	**# of Patients**	**Percentage**
1	**Record failed to succeed**	**67**	**13.16**
0	**No death during the protocol**	**442**	**86.84**

Table 4 Classes at Short Term Analysis

The registered and eligible, according to the AML99 dataset, patients for the study were 509. Several examinations generated 57 different markers for prognosis. Below, there is an analytical explanation of the meaning of each indicator which was directly associated with the classification approach. We use the term “directly associated with” because a small number of the available indicators finally participated to the classification procedure. Due to the variety of them, the indicators were categorized as clinical factors, molecular biology, and cytogenetics abnormalities. Molecular biology is the available information provided by biology at a molecular level. Molecular biology concerns itself with understanding the interactions between the various systems of a cell, including the interactions between DNA, Ribonucleic acid (RNA), and protein biosynthesis and learning how these interactions are regulated **[10]**. On the other hand, cytogenetics is the study of the structure of chromosome material **[11]**. In our case (hematological malignancies), cytogenetics can determine which chromosomal translocations are present in the malignant cells. Finally, clinical variables are all the variables that provide information about the gender, the age, the duration of an illness, and measures from several tests (blood test markers) that do not belong to the cytogenetic or biomoleclar biology studies.

➢ Clinical Factors

Variable	Short Description	Missing Values (# out of 509)	Minimum Value	Maximum Value	Mean Value	Type of measure
Hb_On	Hemoglobin at inclusion	51	3.9	15.9	8.8	gm/dl[1]
PLTS_On	Platelets at inclusion	51	3.0	870.0	52.0	10^9/lt
bl_bm_dia	blasts bone marrow (BM) at diagnosis	0	30.0	99.0	80.0	%
wbc_dia	White blood cell (WBC) at diagnosis	0	0.4	400.0	19.8	10^9/lt
PS_dia	Performance Status from WHO	0	0	3	-	-
cns_on	Central nervous system involved at inclusion	25	0	1	-	-
exm_on	Extra-medullar infiltration at inclusion	33	0	4	-	-
Sex	Male or Female	0	0	1	-	-

Table 5 Clinical Prognostic Factors.

♦ Hb_On

Hemoglobin **[12]**, is the iron-containing oxygen-transport metalloprotein in the red blood cells of the blood in vertebrates and other animals. Hemoglobin transports oxygen from the lungs or gills to the rest of the body, such as to the muscles, where it releases its load of oxygen. Low hemoglobin is referred to as being anemic. Some of the more common reasons are loss of blood (traumatic injury, surgery), nutritional deficiency (iron, vitamin B12), bone marrow problems (replacement of bone marrow by cancer), and abnormal hemoglobin (sickle cell anemia). Higher than normal hemoglobin levels

[1] The hemoglobin level is expressed as the amount of hemoglobin in grams (gm) per decilitre (dl) of whole blood, a decilitre being 100 millilitres.

can be seen in people living at high altitudes and in smokers. The variable given in Table 5 is the measure of hemoglobin at inclusion for every patient.

- **PLTS_On**

Platelets **[13]** are irregularly-shaped, colorless bodies that are present in blood. Their sticky surface lets them, along with other substances, form clots to stop bleeding. A shortage of platelets can cause a person to bleed or bruise easily. Sometimes the number of platelets in the blood is lower than normal. This may be because of illness (including cancer, leukemia or certain blood disorders) or it can be a side effect of chemotherapy treatment. If your bone marrow is not working normally, the number of platelets in your blood may drop. The normal platelet count is between 150 and 400 (10^9/lt). **PLTS_On** variable gives the number of platelets at the inclusion phase (see Table 5).

- **bl_bm_dia**

In AML, the bone marrow makes many unformed cells called blasts **[13]**. Blasts normally develop into white blood cells that fight infection. However, the blasts are abnormal in AML. They do not develop and cannot fight infections. The bone marrow may also make abnormal red blood cells and platelets. The number of abnormal cells (or leukemia cells) grows quickly. They crowd out the normal red blood cells, white blood cells and platelets the body needs. This variable provides information about the number of blasts of a patient at the stage of inclusion, as seen in Table 5.

- **wbc_dia**

White blood cells or leukocytes **[14]** are cells of the immune system which defend the body against both infectious disease and foreign materials. **Wbc_dia** measures the white blast cells of a patient at diagnosis (see Table 5). Several different and diverse types of leukocytes exist. Each has a special role to play in protecting the body against infection. The 3 main types of white blood cells are granulocytes, monocytes, and lymphocytes. Any of the blood-forming cells can turn into a leukemic cell. Once that happens, the cell can reproduce to form many new cancer cells. These cells can overwhelm the bone marrow, spill out into the bloodstream, and spread to other organs.

- **PS_dia**

PS_dia is a categorical variable, scaling from 0 to 3, which classifies the patients according to the characteristics of their AML disease at the diagnostic phase of the threatment (see Table 5). Such variable is the performance status of their disease as scaled by the World Health Organization (WHO) **[15]**. Statistical information about this variable is given by the following table (Table 6). The WHO classification of AML encompasses four major categories:

1. AML with characteristic genetic abnormalities, which includes AML with translocations between chromosome 8 and 21 [t(8;21)], inversions in chromosome 16 [inv(16)], or translocations between chromosome 15 and 17 [t(15;17)]. Patients with AML in this category generally have a high rate of remission and a better prognosis compared to other types of AML.
2. AML with multilineage dysplasia. This category includes patients who have had a prior myelodysplastic syndrome (MDS) or myeloproliferative disease (MPD) that transforms into AML. This category of AML occurs most often in elderly patients and often has a worse prognosis.
3. AML and MDS, therapy-related. This category includes patients who have had prior chemotherapy and/or radiation and subsequently develop AML or MDS. These leukemias may be characterized by specific chromosomal abnormalities, and often carry a worse prognosis.
4. AML not otherwise categorized. Includes subtypes of AML that do not fall into the above categories.

PS at diagnosis			
PS_dia	**WHO scaling**	**# of Patients**	**Percentage**
0	1	235	46.17
1	2	188	36.94
2	3	75	14.73
3	4	11	2.16

Table 6 PS at diagnosis

- **cns_on**

A potential, but uncommon, problem in the treatment of AML is Central Nervous System (CNS) relapse **[16]**. CNS relapse appears to be a particular risk for AML M4 and M5 subtypes. The leukemic cells can enter the Celebro-Spinal Fluid (CSF), a fluid that surrounds the brain and spinal cord, and the drugs do not lead to sufficient accumulation in the fluid. There is therefore a risk that leukemia cells may survive in this site. Variable **cns_on** is a binary variable which classifies the patients, at the stage of inclusion, based on their affection to the CNS involvement (see Table 5). Further information about this variable is given by Table 7.

central nervous system (CNS) involved at inclusion		
cns_on	**# of Patients**	**Percentage**
No	442	98.00
Yes	9	2.00
no value	58	----

Table 7 central nervous system (CNS) involved at inclusion

- **exm_on**

In acute myeloid leukemia (AML), extramedullary disease often occurs as a mass or tissue infiltration **[17]**. Generally referred to as granulocytic sarcomas (GS), but also known as myelosarcoma, myeloblastoma or chloroma (because of the greenish hue of the cut surface) these tumors may occur at various sites of the body, including the skin, bones, orbits, soft tissue, and central nervous system. The extra-medullary infiltration is a categorical variable (see Table 5) that classifies the patients into the following responses.

Extra-medullary infiltration at inclusion		
Exm_on	**# of Patients**	**Percentage**
No infiltration	371	86.28
Lymphnodes	27	6.28
Cutaneous	13	3.02
Both	2	0.47
Other	17	3.95
No value	79	---

Table 8 Extra-medullary infiltration at inclusion

- **Sex**

This binary variable gives information about the gender of the patients.

sex			
Sex	**Short Description**	**Percentage**	**# of Patients**
1	Male	50.88	259
2	Female	49.12	509

Table 9 Gender categorization

- **FAB**

In the 1970s, an international conference of leukemia experts was held to decide on the best system for classifying acute leukemia. The French-American-British (FAB) classification system divided AML into 8 subtypes, M0 through to M7 (see Table 10), based on the type of cell from which the leukemia developed and its degree of maturity **[18]**. This is done by examining the appearance of the malignant cells under light microscopy and/or by using cytogenetics to characterize any underlying chromosomal abnormalities. The subtypes have varying prognoses and responses to therapy. The subtypes indicate degree or lack of maturation of the cells **[18]**, where:

- MO and M1 are characterized by blasts with no or little maturation, including no or little myeloperoxidase reaction, respectively.
- M2 is characterized by some maturation including the presence of promyelocytes or more mature neutrophils. In one form of M2, there is an increase in basophils.
- M3 is characterized by presence of many abnormal promyelocytes and is also referred to as acute promyelocytic leukemia (APL).
- M4 and M5 are characterized by some monocytic differentiation. M4E is a subtype of M4 and is characterized by presence of abnormal eosinophils. M5 is divided into M5A, which is characterized by the presence of many monoblasts and M5B which is characterized by the presence of more mature monocytic cells. M4 is also referred to as acute myelomonocytic leukemia and M5 is also referred to as acute monocytic leukemia.
- M6 is characterized by some erythroid differentiation and is also referred to as erythroleukemia and can be considered a form of di Guglielmo's syndrome.
- M7 is characterized by some megakaryocytic differentiation and is also referred to as acute megakaryoblastic leukemia.

Below (see Table 10), a statistical study indicating the degree of difficulty for prognosis for each subtype **[18]** is outlined. When a patient entered the AML99 protocol, was examined based on the above characteristics and was automatically classified to its belonging group.

Table 11 summarizes all the possible FAB categories at the AML99 protocol. The categorical variable **fab_dia** provides relevant information about the FAB type of a patient before the induction treatment (at inclusion). Note that in this dataset, there was no patient with FAB type M3, which statistically corresponds to a good prognosis.

FAB subtype	Name	Prognosis compared to average for AML
M0	Undifferentiated acute myeloblastic leukemia	Worse
M1	Acute myeloblastic leukemia with minimal maturation	Average
M2	Acute myeloblastic leukemia with maturation	Better
M3	Acute promyelocytic leukemia	Best
M4	Acute myelomonocytic leukemia	Average
M5	Monocytic leukemia	Average
M6	Acute erythroid leukemia	Worse
M7	Acute megakaryoblastic leukemia	Worse

Table 10 FAB type description [18]

FAB at diagnosis		
fab_dia	**# of Patients**	**Percentage**
Missing	**5**	**0.98**
M0	**31**	**6.09**
M1	**90**	**17.68**
M2	**165**	**32.42**
M4	**89**	**17.49**
M4E	**14**	**2.75**
M5A	**66**	**12.97**
M5B	**32**	**6.29**
M6	**15**	**2.95**
M7	**2**	**0.39**

Table 11 French-American-British classification system

➢ Cytogenetic Variables

Before proceeding with the explanation of the several cytogenetic abnormalities from the AML99 protocol it is important to refer to the following terminology.

Analytically, we have four general categories of cytogenetic abnormalities. These are **[19]**:

1. Chromosomal translocation (t) is the process by which a break in at least two different chromosomes occurs, with exchange of genetic material between the chromosomes. Reciprocal translocation refers to an exchange in which there is no obvious overall loss of chromosomal material. An example of a reciprocal translocation is the chromosome t(9;11)(p21;q23) (see below).
2. Chromosomal deletion (del) means loss of chromosomal material. An interstitial deletion results from two breaks in a single chromosome with the loss of intervening material. An example of an interstitial deletion is the 5q- syndrome [del(5q)], in which a variable portion (often the segment between bands q13 and q33) of the long arm of chromosome 5 is lost.
3. Monosomy is a form of genetic loss in which an entire chromosome is lost (eg, monosomy 7 or -7).
4. Chromosomal inversion (inv) requires two breaks in the same chromosome with rotation of the intervening material. An example is inv(16)(p13q22), wherein genes previously on opposite ends of chromosome 16 are juxtaposed after the rearrangement.

Acquired chromosome aberrations are present in the marrow of most patients with acute myeloid leukaemia (AML) at diagnosis **[19]**. Cytogenetically, AML is a very heterogeneous disease with over 160 structural chromosome abnormalities observed recurrently to date. Molecular dissection of many reciprocal translocations and inversions has resulted in cloning of the genes involved in leukaemogenesis. Some recurrent aberrations and the resulting gene rearrangements, namely inv(16)/t(16;16), t(8;21), t(15;17), and rearrangements of band 11q23 and the MLL gene, are now used to help define distinct disease entities within AML. Moreover, cytogenetic abnormalities, whether molecularly characterized or not, are among the most important, independent prognostic factors in AML, and are being used in the management of AML patients.

Treatment of acute myeloid leukemia (AML) depends upon the exact sub-type of leukemia. Acute myeloid leukemia (AML) is classified into seven subtypes M1 to M7 by

the French American British (FAB) classification system. Out of these M3 has got the best prognosis as a targeted therapy called ATRA (all-trans retinoic acid) is available for its treatment. The prognosis in all other AML is based on the cytogenetic analysis with most favorable outcome in translocations (8;21) and (16;16). Patients with deletion in chromosome 5, 7 or 3 have worst prognosis. In general, patients with AML whose cells have translocations seem to fare better than those whose cells have deletions. The poor prognosis associated with increased age may be related to the higher incidence of genetic deletions.

- **inv(16) or t(16;16) [20]**

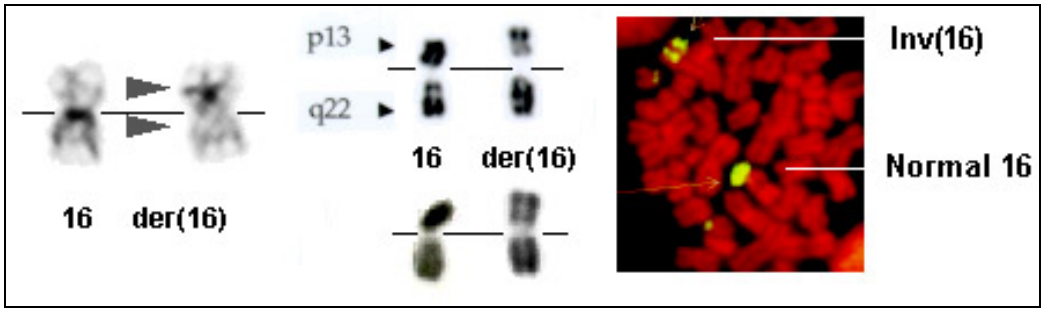

Figure 4 Inversion inv(16) [20]

Patients with inv(16) usually correspond to the subclass of AML M4, with a specific abnormal eosinophil component, considered as a distinct entity in correlation with these specific chromosomal abnormalities. These cases of AML M4 are referred as AML M4EO. In addition to the morphological features of AML M4 excess of monocytes), the bone marrow shows a variable number of eosinophils at all stages of maturation without significant maturation arrest. The most striking abnormalities involve the immature eosinophilic granules. Those are mainly evident at the promyelocyte and myelocyte stages.

- **t(8;21) [20]**

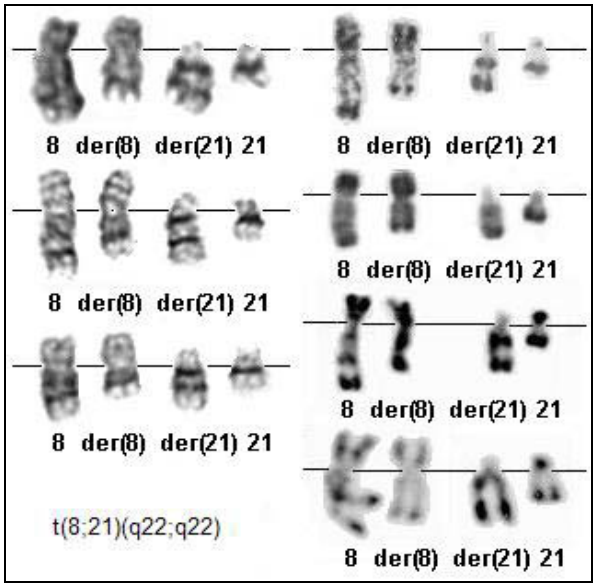

Figure 5 Translocation t(8;21) [20]

Translocation t(8;21) is found in 5-12% of AML. Among the non-random chromosomal aberrations observed in AML, t(8;21)(q22;q22) is one of the best known and usually correlates with AML M2, with well defined and specific morphological features. The common morphological features include the presence of large blast cells with abundant basophilic cytoplasm, often containing numerous azurophilic granulations; few blasts in some cases show very large granules (pseudo-Chediak-Higashi granules), suggesting abnormal fusion. Auer rods are frequently found. In addition to the large blast cells, there are also some smaller blasts, predominantly found in the peripheral blood. Promyelocytes, myelocytes and mature granulocytes with variable dysplasia are seen in the bone marrow.

- **+8 or trisomy 8 [20]**

Trisomy 8 is the most frequently observed trisomy in acute myeloid leukemia (AML) occurring as a sole karyotype abnormality or in addition to other chromosome aberrations. However, despite the high frequency of +8, much remains to be elucidated as regards its epidemiology, etiology, clinical impact, association with other chromosomal abnormalities, cell of origin, and functional and pathogenetic consequences.

- **11q23 [20]**

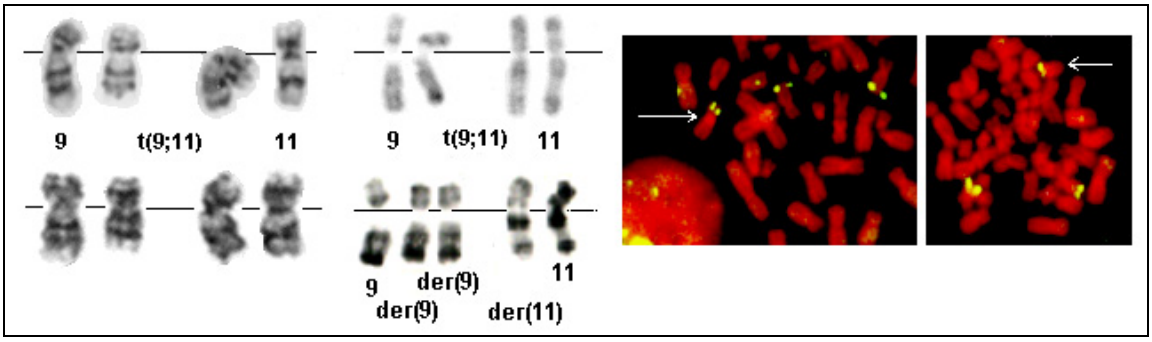

Figure 6 Abnormality 11q23 [20]

Molecular studies have identified a human homologue of the drosophila trithorax gene (MLL). MLL is a developmental regulator and is structurally altered in leukemia associated translocations that show an abnormality at band 11q23. The MLL gene on 11q23 is involved in a number of translocations with different partner chromosomes. The most common translocations observed in childhood AML are the t(9;11)(p21;q23) and the t(11;19)(q23;p13.1); other translocations of 11q23 involve at least 30 different partners chromosomes. Molecular studies have shown that MLL is rearranged more frequently than is revealed by conventional cytogenetic studies. A partial tandem duplication of MLL gene has also been reported in the majority of adult patients whose leukemic blast cells have a +11 and in some with normal karyotype. There is a strong association between AML M5/M4 and deletion and translocations involving 11q23.

- **t(6;9) [20]**

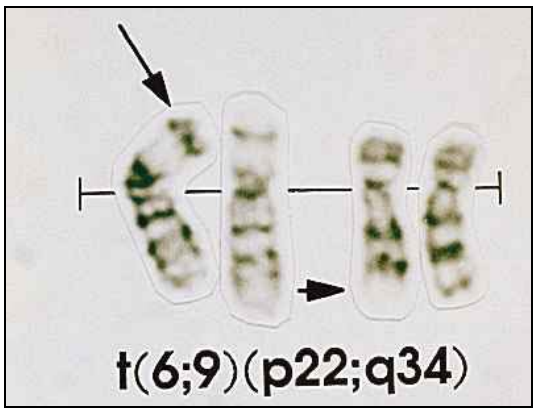

Figure 7 Translocation t(6;9) [20]

The translocation t(6;9) results in the formation of a chimeric fusion gene: DEK (6q23) and CAN (9q34). CAN is a putative oncogene which may be activated by fusion of its 3' end to other genes than DEK.

- **t(9;22) [21]**

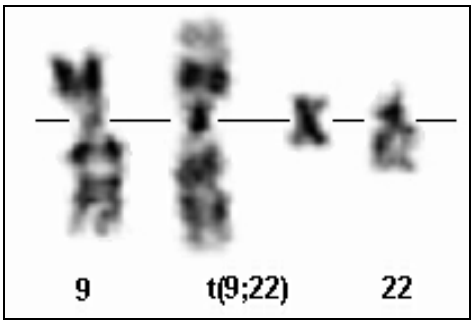

Figure 8 Translocation t(9;22) [21]

Translocation (9;22) is a specific chromosomal abnormality that is associated with chronic myelogenous leukemia (CML). It is due to a reciprocal translocation designated as t(9;22)(q34;q11), which means an exchange of genetic material between region q34 of chromosome 9 and region q11 of chromosome 22. The presence of this translocation is a highly sensitive test for CML, since 95% of people with CML have this abnormality. However, the presence of this chromosome is not sufficient to diagnose CML, since it is also found occasionally in acute myelogenous leukemia (AML).

- **t(15;17)(q22;21) [21]**

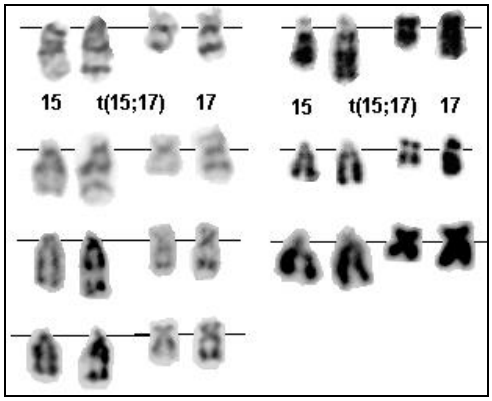

Figure 9 Translocation t(15;17) [21]

Translocation t(15;17)(q22;21) is associated consistently with AML M3. This chromosomal abnormality first appeared to be confined to the characteristic or morphologically typical M3 AML or "hypergranular promyelocytic leukemia", defined by bone marrow replacement with highly granulated blast cells.

- **t(8;16) [21]**

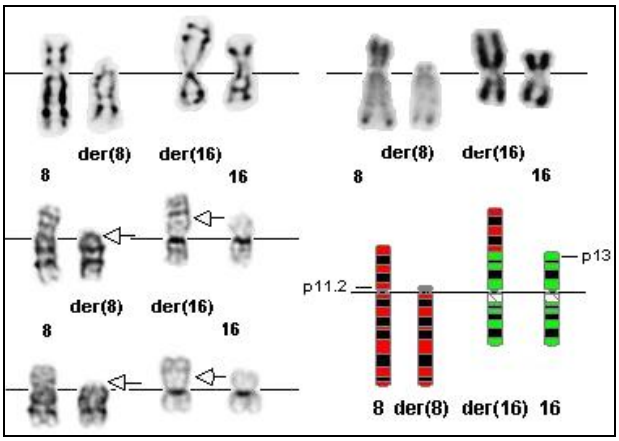

Figure 10 Translocation t(8;16) [21]

The t(8;16) has been cloned and shown to fuse the MOZ (monocytic leukemia zinc finger) gene at 8p11.2 to the CBP (CREB binding protein) gene at 16p13.3. This translocation is associated with AML M5/M4.

- **-5, 12p, -7 [20]**

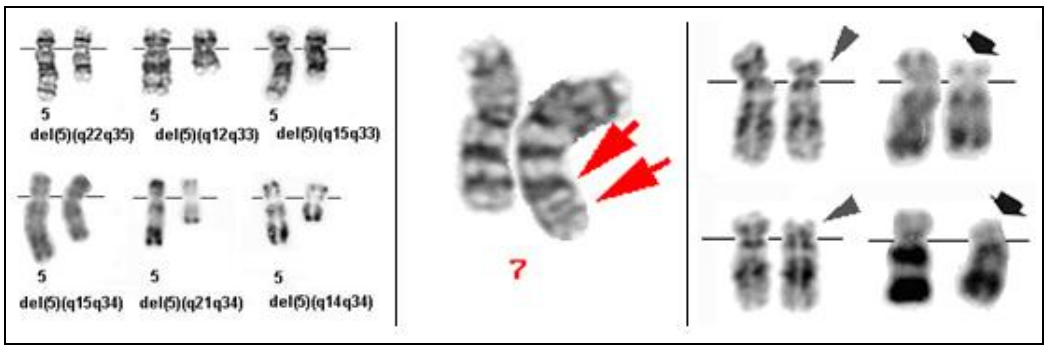

Figure 11 Deletion 5q, deletion 7q and 12p [20]

Del(5q) is an interstitial deletion with variable proximal and distal breakpoints, deletion (7q) correspond to a cluster of breakpoints in 7q11 to 7q36, and 12p abnormalities are common in a broad spectrum of hematological malignancies.

- **inv(3) or t(3,3) [22]**

The t(3;3) is a nonrandom abnormality found in a small percentage of patients with myelodysplastic syndrome, secondary AML or chronic myeloid leukemia and is strongly associated with abnormal thrombopoiesis and a particularly poor prognosis.

Summing up, all the abnormalities described above are summarized into the following table, together with a brief description about their success of prognosis.

Abnormality	FAB-Subgroups	Prognosis
inv(16) or t(16;16)	AML-M4Eo (Myelomonocytic leukemia with eosinophilia)	High CR rate; better prognosis than most other acute non-lymphocytic leukemias.
t(8;21)(q22;q22)	AML-M2	CR in most cases (90%); but relapse is frequent.
+8 or trisomy 8	is present in each FAB subgroup	CR in 60-70% (almost 90% in cases accompanying t(8;21), t(15;17)).
11q23	AML-M4, M5	Very poor in general; variable according to the translocation, the phenotype, the age, and whether the leukemia is de novo or treatment related.
t(6;9)	AML – M1, M2, M4	Remission difficult to obtain; CR in only half cases.
t(9;22)	AML – M1, M2	Very poor
inv(3) or t(3;3)	AML - M1, M2, M4, M6,M7	Very poor
t(15;17)(q21-q11-22)	AML-M3	Good prognosis
t(8;16)(p11;p13)	AML-M4, M5	Good prognosis
-5, 12p, -7	AML – M6, M7	Very poor

Table 12 Categorization of Cytogenetic Variables and Prognostic Group [19]

In case of the AML99 protocol, the above cytogenetic abnormalities are described by variable "**citog**". Note that the cytogenetic information is widely accepted as a good prognostic factor **[3]** in AML disease, and for that reason it was impossible not to be gathered for the purposes of this study. The following table provides statistical information about the possible cytogenetic abnormality of a patient as well as the number of patients who were found with the same abnormality.

Cytogenetics		
Citog	**# of Patients**	**Percentage**
normal kariotype	177	34.77
inv(16)	25	4.91
t(8;21)	37	7.27
+8	21	4.13
t(11)(q23)	8	1.57
t(6;9)	9	1.77
t(9;22)	5	0.98
t(3 ;3)inv(3)	6	1.18
Iperdiploid	19	3.73
complex kariotype	18	3.54
other[2]	65	12.77
Failed	53	10.41
not done	66	12.97

Table 13 Cytogenetic abnormalities and normal kariotype

➢ Molecular Biology Variables

The molecular biology data were clearly asymmetrical: the negative condition represented normality, while the positive one described an abnormal situation. Further information about these indicators is given below. Many biological markers, has been

[2] Alterations : 12p, -7/del(7q), del(9q), t(3 ;5), t(8 ;16), t(15;17), –5/del(5q), del(13q), mark, -5; -7

proved by the biologists to be highly correlated with the cytogenetic abnormalities, and for that reason many of them were excluded from the tested dataset.

- **AML/ETO**

In patients with acute myelogenous leukemia (AML), t(8;21)(q22;q22) is a relatively frequent structural cytogenetic abnormality **[23]**. The AML/ETO fusion protein, created by the (8;21) translocation in M2-type acute myelogenous leukemia (AML), is a dominant repressive form of AML. This effect is due to the ability of the ETO portion of the protein to recruit co-repressors to promoters of AML target genes **[24]**.

AML/ETO[3] (mol.bio.)		
AML_ETO	**# of Patients**	**Percentage**
Neg	388	89.40
Pos	36	8.29
No value	10	2.30

Table 14 Molecular Biology corresponded with translocation t(8; 21)

- **BCR/ABL [25]**

In AML, the translocation occurs between chromosomes 9 and 22 (human DNA is packaged in 23 pairs of chromosomes) and produces a new, abnormal gene called BCR-ABL. This abnormal gene produces Bcr-Abl tyrosine kinase, an abnormal protein that causes the excess White Blast Cells (WBCs) typical of AML.

[3] It is supposed to be a correspondence with cytogenetic variable t(8;21)

BCR/ABL[4] (mol.bio.)		
BCR_ABL	**# of Patients**	**Percentage**
Neg	416	81.73
Pos	6	1.18
No value	87	17.09

Table 15 Molecular Biology corresponded with translocation t(9; 22)

- **DEK/CAN**

DEK-CAN fusion gene is the molecular basis in pathogenesis of AML. The detection of DEK-CAN fusion gene is significant for diagnosis of AML, evaluation of curative effect, and predication of prognosis. Researches has shown that there is a relationship of (6;9) chromosome translocation with DEK-CAN fusion gene expression in patients with acute myeloid leukemia (AML) **[26]**.

DEK/CAN[5] (mol.bio.)		
DEK_CAN	**# of Patients**	**Percentage**
Neg	385	88.71
Pos	9	2.07
No value	40	9.22

Table 16 Molecular Biology corresponded with translocation t(6; 9)

- **inv16 [27]**

Inversion 16 (inv(16)) results in a fusion between genes CBFB on the q arm and MYH11 on the p arm and is associated with acute myeloid leukemia (AML) subtype M4Eo.

[4] It is supposed to be a correspondence with cytogenetic variable t(9;22)
[5] It is supposed to be a correspondence with cytogenetic variable t(6;9)

inv16[6] (mol.bio.)		
inv16	**# of Patients**	**Percentage**
Neg	398	91.71
Pos	25	5.76
No value	11	2.53

Table 17 Molecular Biology corresponded with inversion inv(16)

- **MLL**

Structural abnormality of the 11q23 band (11q23+) bearing the MLL gene translocation (MLL+). MLL is a recurrent chromosome change in leukemia described in acute myeloblastic leukemia (AML) and in acute lymphoblastic leukemia (ALL), with a peak incidence in infant leukemia. 11q23+/MLL+ is described in 3% to 4% of AML cases and is more frequent in younger subjects with de novo AML (5%-7%) **[28]**.

MLL (mol.bio.)		
Mll	**# of Patients**	**Percentage**
Neg	292	57.37
Pos	18	3.54
No value	199	39.10

Table 18 myeloid/lymphoid or mixed-lineage leukemia

The previous molecular biology attributes can be summarized into the following table.

[6] It is supposed to be a correspondence with cytogenetic variable inv(16)

Molecular Biology		
Molec	**# of Patients**	**Percentage**
INV16	25	4.91
AML_ETO	36	7.07
MLL	18	3.54
DEK/CAN	9	1.77
BCR/ABL	6	1.18
No value	415	81.53

Table 19 Summary of Molecular Biology alterations

Conclusively, the molecular biology and their corresponded cytogenetic abnormalities were summarized into an integrated attribute, named as "**citomol**". The molecular biology markers, given by Table 19, have a high correlation with some cytogenetic abnormalities as noticed above. For that reason, such information was merged with the cytogenetic alterations, provided by Table 13, and finally variable **citomol**, concluded all the available information (see Table 20).

Cytogenetic/ Molecular Biology (integrated)		
Citomol	**# of Patients**	**Percentage**
normal kariotype	170	33.40
inv(16)	28	5.50
t(8;21)	43	8.45
+8	19	3.73
t(11)(q23)	20	3.93
t(6;9)	10	1.96
t(9;22)	6	1.18
t(3 ;3)inv(3)	6	1.18
Iperdiploid	15	2.95
complex kariotype	17	3.34
Other	63	12.38
no value	112	22.00

Table 20 Fused information from Molecular Biology and Cytogenetics

Indicators like the FLT3/ITD, FLT3/d835 and NPM, shown respectively in Table 21, Table 22, and Table 23, also provided relevant molecular biology information.

- **FLT3/ITD**

Mutations in the FLT3 gene are the most common genetic alteration found in AML patients **[29]**. Internal tandem duplication (**itd**) mutations arise from duplications of the juxtamembrane portion of the gene and result in constitutive activation of the FLT3 protein. This alteration has been identified in ~20% to 30% of patients with acute myelogenous leukemia and appears to be associated with a worse prognosis. Variable **itd** is a binary variable which categorizes patients who are affected by this alteration and patient who are not.

FLT3/ITD		
itd	**# of Patients**	**Percentage**
Neg	322	63.26
Pos	86	16.90
missing	101	19.84

Table 21 Internal Tandem Duplication (ITD) in the FLT3 gene

- **FLT3/d835**

Another type of FLT3 mutation is mutations at aspartic acid residue 835 (**d835**), which occurs in ~7.0% of acute myelogenous leukemia cases **[27]**. Binary variable **d835** separates patients with this abnormality.

FLT3/d835		
d835	**# of Patients**	**Percentage**
Neg	352	69.16
Pos	22	4.32
missing	135	26.52

Table 22 Mutations at aspartic acid residue 835

- **NPM**

Mutations of **NPM** in **AML** disrupt the **NPM** nucleolar-localization signal, causing accumulation of **NPM** in the cytoplasm. The mutations were prominent in those with a normal karyotype and were frequently associated with FLT3 mutation. NPM protein shuttles between the nuclei and cytoplasm **[27]**.

NPM		
npm2	**# of Patients**	**Percentage**
Neg	195	38.31
Pos	112	22.00
missing	202	39.69

Table 23 Nucleophosmin (NPM) mutation in the FLT3 gene

2.2 Preparation of the Dataset

The preprocessing consisted of revising the dataset by excluding its biased as well as the non-relevant attributes. Afterwards, the datasets' variables were standardized to $mean = 0$ and $\sigma^2 = 1$, normalized with its values varying from -1 to 1, and every patient with at least one missing value in his/her prognostic factors was censored. Our group has also designed and implemented several techniques for filling the missing values, specifically focusing on the iterative expectation maximization algorithm (EM) **[30]**. The EM imputation was applied to several of our previous studies **[31]**, **[32]**. This time, the highly correlated attributes (correlated cytogenetics and biomolecular attributes) and the high percentage of missing values disinclined us for applying such techniques to the preprocessing procedure.

Furthermore, indicators that provided no significant difference between the patients were also excluded from the dataset. Specifically, the binary indicator "**cns_on"** was rejected as a biased factor due to the fact that 98% of its values were equal to 0. Attribute "**FAB"** was also excluded from the classification approach, because it was deemed non-relevant from the doctors at Rome. Conclusively, our dataset consisted of variables from Table 5, Table 20, Table 21, Table 22, and Table 23 and is given by Table 24.

Variable	# of missing values	Type of data
Sex	0	Binary
wbc_dia	0	Numerical
PS_dia	0	Categorical
bl_bm_dia	0	Numerical
Hb_On	51	Numerical
PLTS_On	51	Numerical
Citomol	112	Categorical
exm_on	79	Categorical
itd	101	Binary
npm2	202	Binary
d835	135	Binary

Table 24 Variables used for classification

As noticed above, patients with at least one missing value in their record were censored from the classification procedure. Due to this fact, when we started building the dataset for classification by adding variables from the pool of indicators (see Table 25), there was a significant reduction to the number of the dataset's samples. The more the variables of a dataset, the less the number of patients to be classified. Our group had to balance between these two cases, and the optimal solution was to generate several datasets by making combinations from the available pool of indicators. Despite the exhaustive and time-consuming classification, such approach we believed that was an accurate way for evaluating the significance of a group of indicators regarding to the prediction accuracy, and reducing the loss of information by the elimination of a patient from the tested dataset. These sets are outlined below.

Variables												# of Samples
1	Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	---	---	---	---	359
2	Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	---	---	---	335
3	Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	---	---	289
4	Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835	---	259
5	Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835	npm	197

Table 25 List of datasets for classification

3 Methods of Analysis

This Chapter gives a thorough glance to the technical part of our research, with analytical reviews to the design of the Support Vector Machines (SVMs), the Least Squares Support Vector Machines (LS-SVMs) and the Hidden Space Support Vector Machines (HS-SVMs).

Furthermore, in the following chapters the reader will find a report of the classification fusion approach, a field of study in which our group has already involved with and adopted to many researches. Finally, an alternative feature selection method, based on the wrapper and filtering methods, will be also analytically described.

3.1 Supervised Learning

3.1.1 Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are general and efficient learning machines. The problem of consistency of learning procedure in machine learning is the one where the empirical risk converges uniformly to the actual risk. To obtain a small actual risk, i.e., a good generalization performance, it is necessary to have a right balance between the empirical risk and the capacity of a learning machine. SVMs can do this, so they can obtain a good generalization performance. SVMs have other attractive properties, for example, SVMs have a unique global optimal solution and avoid the curse of dimensionality. The introduction of kernel methods has made SVMs to have nonlinear process ability. Presently, there are many Mercer kernels available such as Gaussian radial basis function kernel, sigmoid kernel, polynomial kernel, spline kernels, and others. These kernels must satisfy Mercer's condition or they must be symmetric and positive semi definite. In this study the widely used polynomial and the radial basis function (RBF) kernels are applied to our approach. Before proceeding with the technical part of the support vector machines, a brief report of them in order to understand their functionality will be very helpful.

3.1.1.1 Hyper-plane Classifiers

Consider the class of hyper-planes:

$$(\mathbf{w}\cdot\mathbf{x})+b=0 \ \mathbf{w},\mathbf{x}\in\Re^N, b\in\Re \tag{1}$$

corresponding to decision functions:

$$f(x)=\mathrm{sgn}((\mathbf{w}\cdot\mathbf{x})+b) \tag{2}$$

Among all hyper-planes separating the data, a unique one exists yielding the maximum margin of separation between the classes **[33]**. This margin concept is a first important step towards understanding the formulation of support vector machines. In Figure 12 an illustrative example is given of a separable problem in a two-dimensional input space. One can see that there exist several hyperplanes that separate the data of the two classes. The main goal in Support Vector Machines is to define the unique hyperplane that separates the data of the two classes and keeps the widest distance between the two classes.

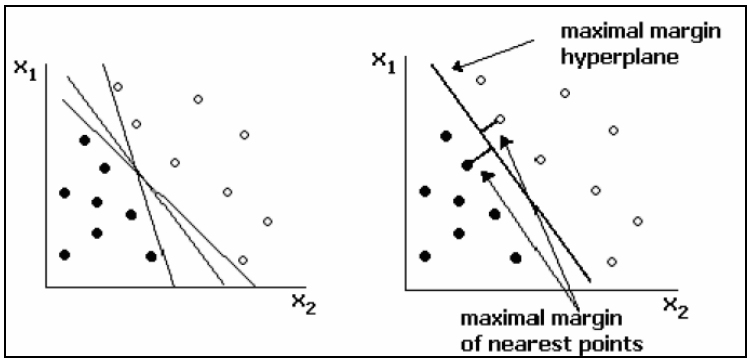

Figure 12 Linear Classification example [33]

As shown in Figure 13, there are two classes to separate $\{1,-1\}$ let y_i denote the label (class) of example x_i such that $y_i\in\{1,-1\}$. The optimal hyper-plane is orthogonal to the shortest line connecting the convex hulls of the two classes, and intersects it half way between the two classes. The problem is separable so there exists a weight vector **w** and a threshold *b* such that $y_i\cdot((\mathbf{w}\cdot\mathbf{x}_i)+b)>0,(i=1,\ldots,n)$. Re-scaling **w** and *b* such that the point(s) closest to the hyper-plane satisfy $\left|(\mathbf{w}\cdot\mathbf{x}_i)+b\right|>0$, we obtain a canonical form (**w**, *b*) of the hyper-plane, satisfying $y_i\cdot((\mathbf{w}\cdot\mathbf{x}_i)+b)\geq 1$. In this case, the *margin*,

measured perpendicularly to the hyper-plane, equals to $2/\|\mathbf{w}\|$. This can be seen by considering two points x_1 and x_2 on opposite sides of the margin, i.e., $(\mathbf{w}\cdot\mathbf{x}_1)+b=1,\ (\mathbf{w}\cdot\mathbf{x}_2)+b=-1$ and projecting them onto the hyper-plane normal vector $\mathbf{w}/\|\mathbf{w}\|$. Thus, we want to maximize $2/\|\mathbf{w}\|$.

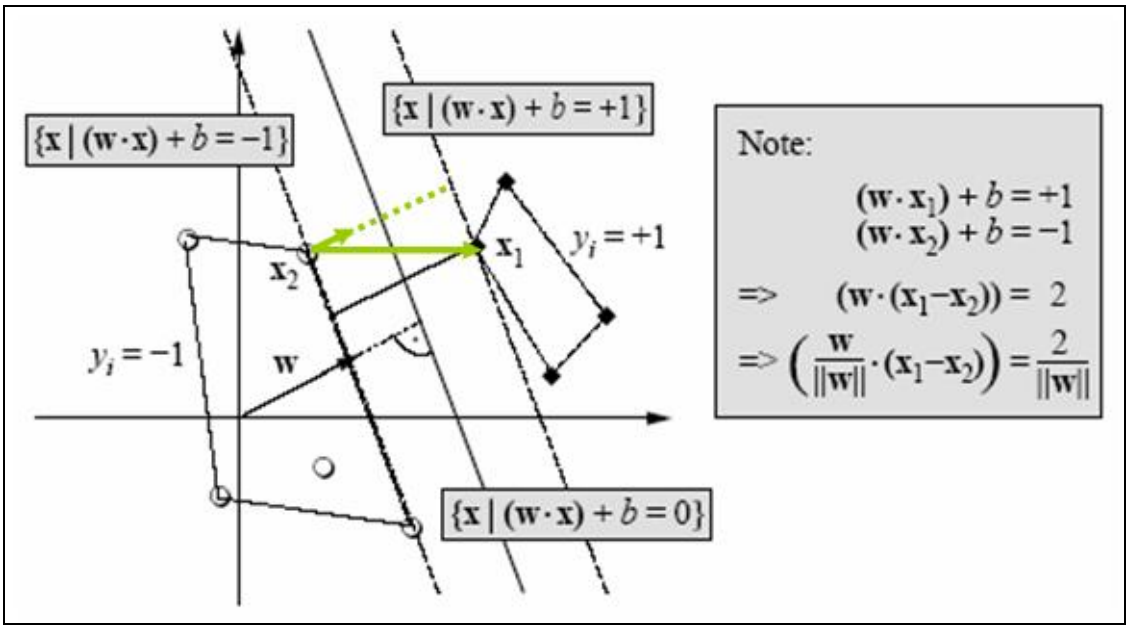

Figure 13 A binary classification toy problem [33]

The above can be formulated in the following form:

$$\min_{w,b}\left(\frac{1}{2}\|\mathbf{w}\|^2\right),\ \textit{subject to}\ y_i\cdot((\mathbf{w}\cdot\mathbf{x}_i)+b)\geq 1,\ i=1,\ldots,N \tag{3}$$

Note that $\max\left(\frac{2}{\|\mathbf{w}\|}\right)=\max\left(\frac{1}{\frac{1}{2}\|\mathbf{w}\|}\right)=\min\left(\frac{1}{2}\|\mathbf{w}\|\right)$. To construct then the optimal hyper-plane one has to solve problem (3). A way to solve the problem is through its Lagrangian dual:

$$\max_{\alpha\geq 0}\left(\min_{w,b}(L(\mathbf{w},b,\alpha))\right) \tag{4}$$

where

$$L(\mathbf{w},b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{i=1}^{N}\alpha_i\cdot(y_i\cdot((\mathbf{w}\cdot\mathbf{x}_i)+b)-1) \tag{5}$$

The Lagrangian (L) has to be minimized with respect to the *primal variables* **w** and *b* and maximized with respect to the *dual variables* α_i. Problem (3) is referred to as the primal problem while problem (5) is referred to as the dual problem. It has been shown that the two problems have the same optimal solution. Therefore we can instead solve the dual which may be an easier problem than the primal. To simplify the dual as $L(\mathbf{w},b,\alpha)$ is convex when α is fixed we proceed as follows:

$$\nabla L(\mathbf{w},b,\boldsymbol{\alpha})=0 \Rightarrow \left\{\begin{array}{l}\dfrac{\partial}{\partial b}L(\mathbf{w},b,\boldsymbol{\alpha})=0\\ \dfrac{\partial}{\partial \mathbf{w}}L(\mathbf{w},b,\boldsymbol{\alpha})=0\end{array}\right\}\Rightarrow\left\{\begin{array}{l}\displaystyle\sum_{i=1}^{m}\alpha_i\cdot y_i=0\\ \mathbf{w}=\displaystyle\sum_{i=1}^{N}\alpha_i\cdot y_i\cdot\mathbf{x}_i\end{array}\right\} \tag{6}$$

Substituting ***w*** into (4) and taking into account the expression $\sum_{i=1}^{m}\alpha_i\cdot y_i=0$ the dual problem can now be written as:

$$\max_{\alpha\in\Re^m}\left(\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\cdot\alpha_j\cdot y_i\cdot y_j(\mathbf{x}_i\cdot\mathbf{x}_j)\right),\ \textit{subject to}\ \left\{\begin{array}{l}\displaystyle\sum_{i=1}^{N}\alpha_i\cdot y_i=0\\ \alpha_i\geq 0,\ i=1,\ldots m\end{array}\right\} \tag{7}$$

Problem (7) can be written in matrix form as follows:

$$\max_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T\cdot\mathbf{1}-\frac{1}{2}\cdot\boldsymbol{\alpha}^T\cdot Q\cdot\boldsymbol{\alpha}\right),\ \textit{with}\ Q=\left[y_i\cdot y_j\cdot(\mathbf{x}_i\cdot\mathbf{x}_j)\right],\ \textit{subject to}\left\{\begin{array}{l}\boldsymbol{\alpha}^T\cdot\mathbf{y}=\mathbf{0}\\ \boldsymbol{\alpha}\geq 0\end{array}\right\} \tag{8}$$

which is a quadratic optimization problem with Q being a positive semi-definite matrix or a positive definite matrix, and can be solved with any quadratic optimization solver. In the case that the matrix is positive definite (all eigenvalues strictly positive), the solution **α** to this QP problem is global and unique. On the other hand, when the matrix is positive semi-definite (all eigenvalues positive but zero eigenvalues possible), the solution is still global but might not be necessarily unique.

3.1.1.2 Support Vectors

From (6) is clear that the solution vector $\mathbf{w}$ is an expansion of the training patterns whose α_i is non-zero. Those patterns are called *Support Vectors*. These vectors in deed lie on the margin because using duality theory and KKT conditions it turns out that these vectors must satisfy the following equation:

$$\alpha_i \cdot \left[y_i \cdot ((\mathbf{w} \cdot \mathbf{x}_i) + b) - 1 \right] = 0, \; i = 1, \ldots N \tag{9}$$

which is called the *complementary slackness condition*. This equation is used to find the value of *b*.

3.1.1.3 Soft Margin Support Vector Classifiers

In practice, a separating hyperplane may not exist, e.g. if a high level of noise causes a large overlap between the classes. Consider Figure 14, where the data from the two classes are not separable.

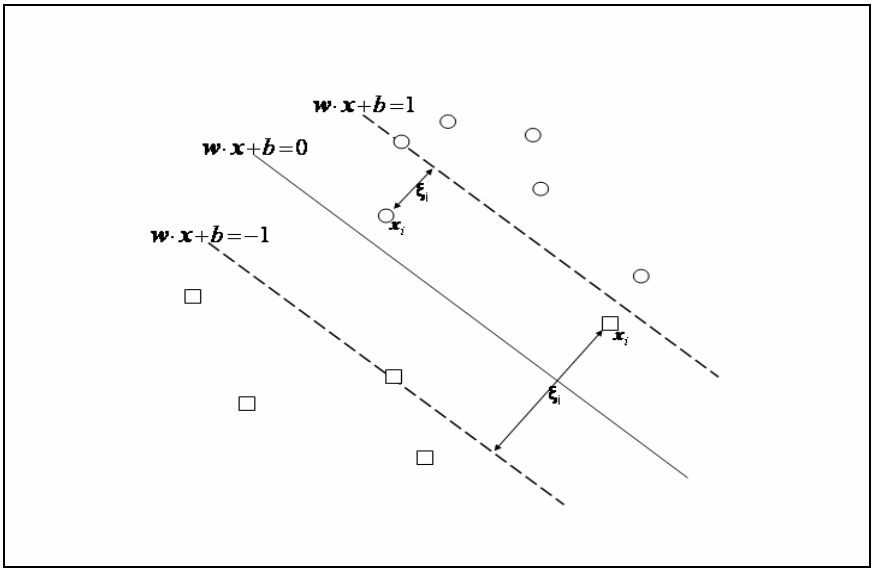

Figure 14 Overlapping classes [33]

The extension of linear SVMs to the non-separable case was made by Cortes and Vapnik **[34]**. Basically, it is done by taking additional slack variables in the problem

formulation. In order to tolerate misclassifications due to the fact that a complete separation of the data might not achieved, one modifies the set of inequalities into:

$$\begin{cases} y_i = 1,\ \mathbf{w}\cdot\mathbf{x}_i + b < 1 \ \Rightarrow \mathbf{w}\cdot\mathbf{x}_i + b + \xi_i \geq 1 \Leftrightarrow \mathbf{w}\cdot\mathbf{x}_i + b \geq 1 - \xi_i \\ y_i = -1,\ \mathbf{w}\cdot\mathbf{x}_i + b > -1 \ \Rightarrow \mathbf{w}\cdot\mathbf{x}_i + b - \xi_i \leq -1 \Leftrightarrow \mathbf{w}\cdot\mathbf{x}_i + b \leq -1 + \xi_i \end{cases} \Rightarrow \tag{10}$$

$$\Rightarrow y_i\left(\left(\mathbf{w}\cdot\mathbf{x}_i\right) + b\right) \geq 1 - \xi_i,\ \xi_i \geq 0,\ i = 1, 2, \ldots, N \tag{11}$$

where ξ_i are just slack variables in optimization theory. Taking the above into consideration we can re-formulate the problem as follows, where C is a positive real constant:

$$\min_{w,b}\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\cdot\sum_{i=1}^{N}\xi_i\right),\ subject\ to \begin{cases} y_i\left(\left(\mathbf{w}\cdot\mathbf{x}_i\right) + b\right) \geq 1 - \xi_i \\ \xi_i \geq 0,\ i = 1, 2, \ldots, N \end{cases} \tag{12}$$

Using duality theory as we did in the original problem and taking partial derivatives the dual problem takes the following form:

$$\max_{\boldsymbol{\alpha}}\left(\boldsymbol{\alpha}^T\cdot 1 - \frac{1}{2}\cdot\boldsymbol{\alpha}^T\cdot Q\cdot\boldsymbol{\alpha}\right),\ Q = \left[y_i\cdot y_j\cdot\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)\right] \tag{13}$$

$$subject\ to \begin{cases} \boldsymbol{\alpha}^T\cdot\mathbf{y} = 0 \\ \mathbf{0} \leq \boldsymbol{\alpha} \leq \mathbf{C} \end{cases} \tag{14}$$

In comparison with the linearly separable case this problem has additional box constraints.

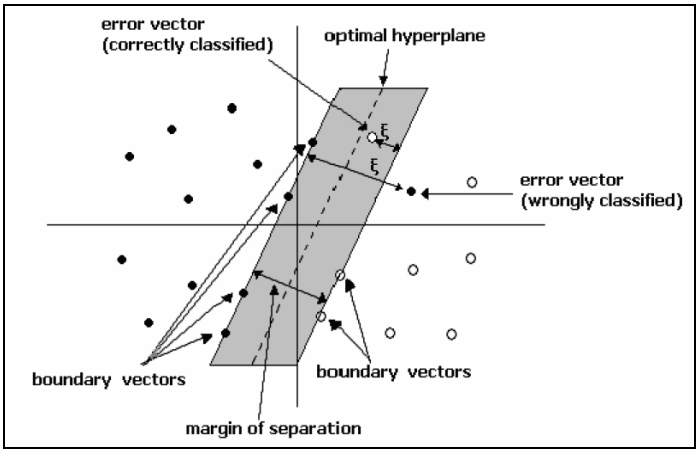

Figure 15 Slack Variables ξ_i [33]

3.1.1.4 Support Vector Machines – Linear Kernels

As was discussed in the previous section Support vector machines try to find an optimum decision boundary of the form:

$$(\mathbf{w} \cdot \mathbf{x}) + b = 0 \tag{15}$$

Using convex optimization and duality theory we derive expressions for ***w*** and *b*:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \mathbf{x}_i \tag{16}$$

$$y_i \cdot ((\mathbf{w} \cdot \mathbf{x}_i) + b) - 1 = 0 \tag{17}$$

where *N* is the number of training samples, $\alpha_i \geq 0$ are Lagrange multipliers, x_i is the ith input sample and y_i is the class label of ith input sample. Following the above, the hyper-plane decision function can be written as:

$$f(x) = \text{sgn}((\mathbf{w} \cdot \mathbf{x}) + b) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot ((\mathbf{x} \cdot \mathbf{x}_i) + b) \right) \tag{18}$$

We can obtain decision functions of the more general form:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \left(\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i) \right) + b \right) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot K(\mathbf{x}, \mathbf{x}_i) + b \right) \tag{19}$$

where *K* is called a kernel.

3.1.1.5 Support Vector Machines – Polynomial Kernels

Using the following transformations:

$$\Phi\left(\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \right) :\mapsto \left[1, \sqrt{2} \cdot x_1, \sqrt{2} \cdot x_2, \cdot x_1^{\,2}, x_2^{\,2}, \sqrt{2} \cdot x_1 \cdot x_2 \right]^T \tag{20}$$

$$\Phi\left(\mathbf{x}_1 = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right) :\mapsto \left[1, \sqrt{2} \cdot y_1, \sqrt{2} \cdot y_2, \cdot y_1^{\,2}, y_2^{\,2}, \sqrt{2} \cdot y_1 \cdot y_2 \right]^T \tag{21}$$

$$\left(\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_1) \right) = \left(1 + x_1 \cdot y_1 + x_2 \cdot y_2 \right)^2 = \left(1 + (\mathbf{x} \cdot \mathbf{x}_1) \right)^2 \tag{22}$$

we can have a polynomial decision function of degree two:

$$f(x) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \left(1 + (\mathbf{x} \cdot \mathbf{x}_i) \right)^2 + b \right) \tag{23}$$

Or generally, polynomial decision function of degree *d*:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \left(1 + (\mathbf{x} \cdot \mathbf{x}_i)\right)^d + b \right) \tag{24}$$

3.1.1.6 Support Vector Machines – Radial Basis Function Kernels

A popular choice of kernel is the Gaussian radial basis function:

$$K\left(\mathbf{x}, \mathbf{x}_1\right) = \exp\left(-\frac{\left\| \mathbf{x} - \mathbf{x}_1 \right\|}{2 \cdot p^2} \right),\ p \in \Re \tag{25}$$

3.1.2 Least Squares Support Vector Machines (LS-SVMs)

Support Vector Machines is a powerful methodology for solving problems in nonlinear classification, function estimation and density estimation which has also led recently to many new developments in kernel based learning in general. In these methods one solves convex optimization problems, typically quadratic programs. We focus on Least Squares Support Vector Machines **[35]** which are reformulation to standard SVMs.

Due to the equality constraints in the formulation, a set of linear equations has to be solved instead of a quadratic programming problem. Having the same decision function as in case of Support Decision Machines

$$y\left(x\right) = \operatorname{sgn}\left(\mathbf{w} \cdot \phi\left(\mathbf{x}\right) + b\right) \tag{26}$$

or in case of linear SVM classification

$$y\left(x\right) = \operatorname{sgn}\left(\mathbf{w} \cdot \mathbf{x} + b\right) \tag{27}$$

Suykens proposed the following SVM modification [10]:

$$\begin{aligned} &\min_{w,b,e} J_P\left(w,e\right) = \frac{1}{2}\left\|w\right\|^2 + \gamma \frac{1}{2} \sum_{i=1}^{N} e_i^2 \\ &\textit{subject to } y_i\left(w \cdot \phi\left(x_i\right) + b\right) = 1 - e_i,\ i = 1, 2, \ldots, N \end{aligned} \tag{28}$$

where $\phi(\bullet)$ is the mapping to the high dimensional feature space as in the standard Support Vector Machine case and γ is a constant variable that plays similar role with constant variable C.

The Vapnik formulation is modified here at two points. First, instead of inequality constraints one takes equality constraints where the value 1 at the right hand side is rather considered as a target value than a threshold value. Upon this target value an error variable e_i is allowed such that misclassifications can be tolerated in the case of overlapping distributions. These error variables play a similar role as the slack variables ξ_i in SVM formulations. Second, a squared loss function is taken for this error variable.

The Lagrangian for the problem is:

$$L(w,b,e,\alpha) = J_P(w,e) - \sum_{i=1}^{N} \alpha_i \left(y_i \left(w \cdot \phi(x) + b \right) - 1 + e_i \right) \tag{29}$$

where the α_i values are the Lagrange multipliers, which can be positive or negative now due to the equality constraints. The conditions for optimality yield:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{i=1}^{N} \alpha_i \cdot y_i \cdot \phi(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow \sum_{i=1}^{N} \alpha_i \cdot y_i = 0 \\ \dfrac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma \cdot e_i, \ i = 1,2,\ldots,N \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \rightarrow y_i \left(w \cdot \phi(x) + b \right) - 1 + e_i = 0, \ i = 1,2,\ldots,N \end{cases} \tag{30}$$

The classifier in the dual space takes the form:

$$y(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} a_i \cdot y_i \cdot K(x, x_i) + b \right) \tag{31}$$

similar to the standard Support Vector Machine case.

3.1.3 Hidden Space Support Vector Machines (HS-SVMs)

In Hidden Space Support Vector Machines (HS-SVMs) **[36]**, the input patterns are mapped into a high-dimensional hidden space by a set of hidden nonlinear functions and then the structural risk is introduced into the hidden space to construct HS-SVMs. Moreover, the conditions for the nonlinear kernel function in HS-SVMs are more relaxed. Compared with support vector machines (SVMs), HS-SVMs can adopt more kinds of kernel functions because the positive definite property of the kernel function is not a necessary condition. HS-SVMs, require a little more complex implementation, where two layers of Kernels are used instead of one in SVMs.

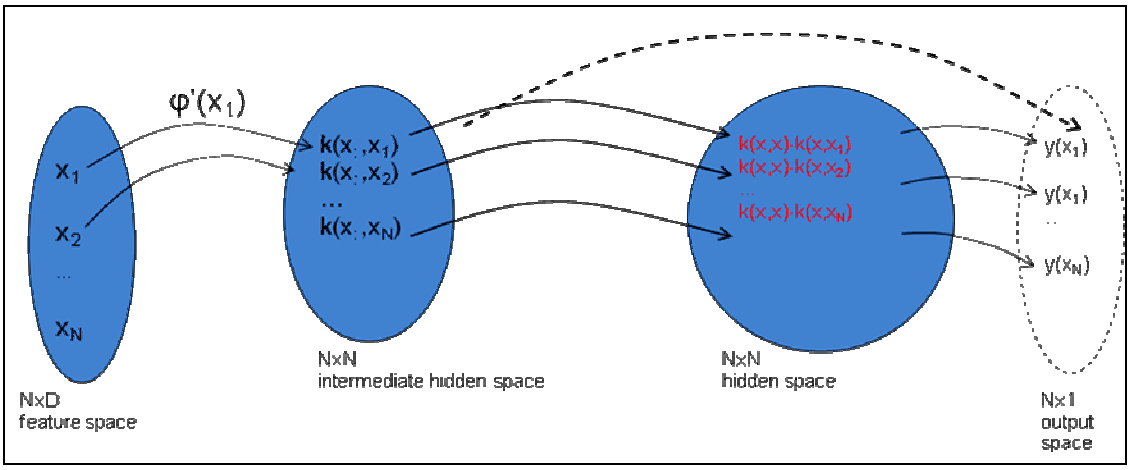

Figure 16 Graphical presentation of a two-layer Kernel

3.1.3.1 Hidden Space

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ denote the set of N independently and identical distributed patterns. Define a vector made up of a set of real valued functions $\Phi_i(\mathbf{x}) \mid i = 1, 2, \ldots, d_1$, as shown by

$$\Phi(\mathbf{x}) = \left[\Phi_1(\mathbf{x}), \Phi_2(\mathbf{x}), \ldots, \Phi_{d_1}(\mathbf{x})\right]^T \tag{32}$$

Since the set of functions $\{\Phi_i(\mathbf{x})\}$ play a role similar to that of a hidden unit in Forward Neural Networks (FNNs), $\Phi_i(\mathbf{x}) \mid i = 1, 2, \ldots, d_1$ are referred to as hidden functions.

Accordingly, the space $Z=\left\{z \mid z=\left[\Phi_1(\mathbf{x}),\Phi_2(\mathbf{x}),\ldots,\Phi_{d_1}(\mathbf{x})\right]^T,\mathbf{x}\in X\right\}$ is called the hidden space or feature space. Now consider a special kind of hidden function: the real symmetric kernel function $K(\mathbf{x},\mathbf{y})=K(\mathbf{y},\mathbf{x})$. Let $X=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ and the kernel mapping be

$$\mathbf{x}\xrightarrow{K}z=\left[K(\mathbf{x}_1,\mathbf{x}),K(\mathbf{x}_2,\mathbf{x}),\ldots,K(\mathbf{x}_N,\mathbf{x})\right]^T \tag{33}$$

The corresponding hidden space on *X* can be expressed as

$$Z=\left\{z \mid z=\left[K(\mathbf{x}_1,\mathbf{x}),K(\mathbf{x}_2,\mathbf{x}),\ldots,K(\mathbf{x}_N,\mathbf{x})\right]^T,\mathbf{x}\in X\right\} \tag{34}$$

whose dimension is *N*. It is only the symmetry of kernel functions that is required, which will extend the use of usable kernels in HS-SVMs while mercer's condition is required in SVMs. Some usual hidden functions are given:

- **Sigmoid Kernel**

$$K(\mathbf{x},\mathbf{x}_1)=\tanh\left(p_1\cdot\mathbf{x}^T\cdot\mathbf{x}_1+p_2\right),\ p_1,p_2\in\Re \tag{35}$$

Usually the sigmoid kernel is not positive definite, which limits its application in SVMs. But it is not a problem in HS-SVMs. If parameter $p_1\to\infty$ in the sigmoid kernel, we will obtain a sign function.

$$K(\mathbf{x},\mathbf{x}_1)=\operatorname{sgn}\left(p_1\cdot\mathbf{x}^T\cdot\mathbf{x}_1+p_2\right),\ p_1,p_2\in\Re \tag{36}$$

Although the differential of the sign function does not exist at some points, it also can be used in HS-SVMs.

- **Gaussian Radial Basis Kernel**

$$K(\mathbf{x},\mathbf{x}_1)=\exp\left(-\frac{\|\mathbf{x}-\mathbf{x}_1\|}{2\cdot p^2}\right),\ p\in\Re \tag{37}$$

which is a widely used Kernel.

- **Polynomial Kernel**

$$K(\mathbf{x},\mathbf{x}_1)=\left(\mathbf{x}^T\cdot\mathbf{x}_1+1\right)^d,\ \mathrm{d}\in\mathbb{N} \tag{38}$$

which is a positive definite kernel used in SVMs frequently.

- **Generalized Multiquadratics Kernel**

$$K(\mathbf{x},\mathbf{x}_1)=\left(1+\left\|\mathbf{x}-\mathbf{x}_1\right\|^2\right)^d,\ \mathrm{d}\notin\mathbb{Z} \tag{39}$$

If $d<0$ is a mercer admissible kernel.

3.1.3.2 HS-SVMs for Pattern Recognition

Let a pattern set be $X=\left\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_N,y_N)\,|\,\mathbf{x}_i\in\Re^d,y_i\in\{-1,1\}\right\}$ and a kernel function be $K(\mathbf{x},\mathbf{x_i})$. The mapped patterns in the hidden space *Z* can be expressed as $\left\{(\mathbf{z}_1,y_1),(\mathbf{z}_2,y_2),\ldots,(\mathbf{z}_N,y_N)\,|\,\mathbf{z}_i=\left[K(\mathbf{x}_1,\mathbf{x}_i),K(\mathbf{x}_2,\mathbf{x}_i),\ldots,K(\mathbf{x}_N,\mathbf{x}_i)\right]^T\right\}$. The training procedure is to solve the dual problem. In the process of transforming the primal problem to the dual one we derive an expression for ***w***:

$$\mathbf{w}=\sum_{i=1}^{N}\alpha_i\cdot y_i\cdot\mathbf{z}_i=\sum_{i=1}^{N}\alpha_i\cdot y_i\cdot\left[K(\mathbf{x}_1,\mathbf{x}_i),K(\mathbf{x}_2,\mathbf{x}_i),\ldots,K(\mathbf{x}_N,\mathbf{x}_i)\right]^T \tag{40}$$

The threshold value *b* can be obtained in a similar way to SVMs. The decision function of HS-SVMs takes the following form:

$$y=\mathrm{sgn}\left((\mathbf{w}\cdot\mathbf{x})+b\right)=\mathrm{sgn}\left[\sum_{i,j=1}^{N}\alpha_i\cdot y_i\cdot K(\mathbf{x}_i,\mathbf{x}_j)\cdot K(\mathbf{x}_j,\mathbf{x})+b\right] \tag{41}$$

3.1.3.3 Proposal of Projection Pursuit SVM

From the theory of SVM, see chapter 3.1.1, we know that a linear decision function can be written in the form:

$$f(x)=\mathrm{sgn}((\mathbf{w}\cdot\mathbf{x})+b)=\mathrm{sgn}\left(\sum_{i=1}^{N}\alpha_i\cdot y_i\cdot((\mathbf{x}\cdot\mathbf{x}_i)+b)\right) \tag{42}$$

or in the more general kernel form:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} \alpha_i \cdot y_i \cdot K(\mathbf{x}, \mathbf{x}_i) + b \right) \tag{43}$$

On the other hand Hidden Space SVM arrives, as shown in chapter 3.1.3.2, to a decision function:

$$f\left(\mathbf{x}\right) = \operatorname{sgn}\left[\sum_{i,j=1}^{N} \alpha_i \cdot y_i \cdot K\left(\mathbf{x}_i, \mathbf{x}_j\right) \cdot K\left(\mathbf{x}_j, \mathbf{x}\right) + b \right] \tag{44}$$

Let $X = \left\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\right\}$ be the training set. If we define a mapping $X \to Z$ of the form $Z = K'\left(X\right)$ then we may find a new space Z where X (the training set) may be more independent and linearly separable. This new vector space Z could be described by functions of the form:

$$\mathbf{z}_i = K'\left(\mathbf{x}_i\right), i = 1, 2, \ldots N \tag{45}$$

then equation (44) using (42) could be written as:

$$f\left(\mathbf{x}\right) = \operatorname{sgn}\left[\sum_{i,j=1}^{N} \alpha_i \cdot y_i \cdot K\left(K'\left(\mathbf{x}\right) \cdot K'\left(\mathbf{x}_i\right)\right) + b \right] = \operatorname{sgn}\left[\sum_{i,j=1}^{N} \alpha_i \cdot y_i \cdot K\left(\mathbf{z} \cdot \mathbf{z}_i\right) + b \right] \tag{46}$$

or finally as:

$$f\left(\mathbf{x}\right) = \operatorname{sgn}\left[\sum_{i,j=1}^{N} \alpha_i \cdot y_i \cdot K\left(\mathbf{z}, \mathbf{z}_i\right) + b \right] \tag{47}$$

The transformation described in (45) could be function driven, $K'\left(., X\right)$ like in HS-SVM or data driven $K'_X\left(.\right)$, as in Linear Dependent Analysis (LDA), Principal Component Analysis (PCA), Independent Component Analysis (ICA), etc, where $K'_X\left(.\right)$ is a linear operator with a matrix K'_X defined by the training space.

Consequently, the Hidden Space SVM is a reliable tool for a good separation of the classes because the range of nonlinear mapping (kernel) functions used in HSSVMs becomes larger than that for the traditional SVMs. In other words, the HS-SVMs can adopt more kinds of kernel functions because the positive definite property of the kernel function is not a necessary condition. Moreover, HS-SVMs have the same computation complexity as SVMs. The proof of this theory is given in **[36]**.

3.1.4 Classifier Fusion

The field of study presented in this chapter is an extension of our work published in **[31]**. Classifier ensembles have in the recent years produced promising results, improving accuracy, confidence and most importantly feature space coverage in many practical applications. The recent trend is to move from heuristic combinations of classifiers to more statistically sound integrated schemes and produce quantifiable results as far as error bounds and overall generalization capability are concerned. Multi-classifier systems have emerged from the realization that we can provide better results using a collection of even relatively poor-performing elementary classifiers, than by utilizing a single fine-tuned one. The perquisites to this include a good selection strategy and homogenization of each classifiers inputs and outputs.

Note that an improvement on the single best classifier or on the group's average performance, for the general case, is not guaranteed. What is exposed here are only "clever heuristics." However, the experimental work published so far and the theories developed for a number of special cases demonstrate the success of classifier combination methods.

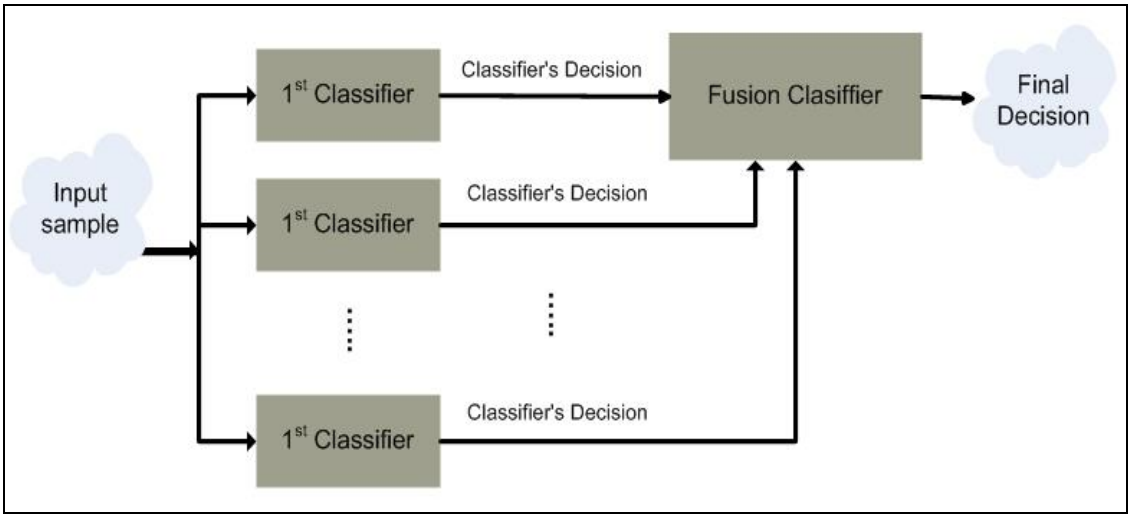

Figure 17 Classifier Fusion

3.1.4.1 Classifier Fusion Background

In combining classifiers the key objective was to obtain the most accurate feature mapping by maintaining diversity and simplicity. There are various approaches in related literature as to the available fusion techniques **[37]-[45]**. They range from simple majority voting and averaging combiners to Bayesian probabilistic models and hyper-classifiers that work on the feature space composed by the soft outputs of the individual classifiers. There is also significant recent research effort in providing statistical foundations for the existing combiners **[45]**. Another trend is to select and use different features for each classifier as in **[46]**.

A major discrimination between the various approaches is based on the type of input data used. Most early classifier combiners used the crisp labels of the individual classifiers. They relied on schemes such as majority voting which extract posterior class probability statistics though counting the true and assigned labels per class. Schemes based on this approach include Behavior Knowledge Space method, naïve Bayes combination and simple or weighted majority voting.

Another group of classifier fusion methods utilizes the soft outputs of the individual classifiers operating on the dataset's features. Having available continuous-valued outputs, i.e. more information per sample to work with, these algorithms are in theory more effective. The fuzzy support values can represent probabilities and even convey information on the confidence that each specific learner places on its class estimate. Some of the simpler schemes belonging to this group include min, max, average and product combiners. Finally, some combiners do not need training after the classifiers in the ensemble have been trained individually, and other combiners need additional training.

At a more advanced level one can apply probabilistic product, linear combiners, linear, quadratic and Fisher discriminant functions. State of the art research in this area though focuses on using the decision profile (*DP*) to calculate Decision Templates (*DTs*) **[47]** and Dempster-Shafer membership degrees for each sample. Classical experts like neural networks, logistic classifiers and other linear or nonlinear classifiers can also be used. The later approach however requires reshaping (unfolding) the DP to form a new output feature space. Additional information about the classifier ensembles and the

reasons for discriminating them into several groups will be given in the following chapters.

3.1.4.2 Diversity Measures

The success of classifier fusion depends on two factors, a pool of diverse individual classifiers to be fused and the proper combining method. If two classifiers in a three classifier fusion task are completely redundant, many fusion schemes will exhibit poorer performance. If two classifiers agree everywhere, the fusion will not achieve any accuracy improvement no matter what fusion method is used. For this reason measures that assess the diversity among the group of classifiers to be fused play a significant role in the entire process.

There are different diversity measures available from different fields of research **[44]**. Some of these measures, such as the Q-statistic and the correlation coefficient have come directly from mainstream statistics whilst others have developed through the field of statistical pattern recognition, specifically for the problems of multiple classifier systems. Some of these measures work on the whole group of L classifiers whilst other measures consider the classifiers on a pairwise basis and then average the results. We can also consider the measures of diversity into two groups:

- Measures looking for diversity; the higher the value the more diverse $(\uparrow)$.
- Measures looking for similarity: the higher the value the less diverse $(\downarrow)$.

The diversity measures used in this study are summarized into the following table. The way that they estimate the correlation degree in a group of classifiers is given by the following example. Assume that we have a set of M Level 1 classifiers, with $M > 2$, and we want to assess the correlation degree of this group. A non-pairwise technique straightforward evaluates the diversity of this group by a single value. On the other hand, a pairwise measure respectively calculates the diversity degree of all the possible pairs of

classifiers from this group. Finally, the overall correlation degree of the group is the average value of the $\binom{M}{2} = \frac{M!}{(M-2)! \cdot 2!}$ measures.

Measure of Diversity	Belonging group	
The Q statistic (Q)	(↓)	Pairwise
The Correlation Coefficient (p)	(↓)	Pairwise
Disagreement Measure (D)	(↑)	Pairwise
The Double-Fault Measure (DF)	(↓)	Pairwise
Kappa Statistic (k)	(↓)	Pairwise
The Entropy Measure (E)	(↑)	Non-Pairwise
Kohavi-Wolpert variance (kw)	(↑)	Non-Pairwise

Table 26 Measures of Diversity

3.1.4.2.1 Pairwise Diversity Measures

Before we proceed with the presentation of the pairwise measures, it is essential to refer to the relationship table between two classifiers. This table provides significant information about the probabilities for the respective pair of correct/incorrect outputs. In other words, if we have two classifiers, D_i and D_j, then a 2×2 relationship table with probabilities that summarises their inputs is produced (see Table 27). Note that $a + b + c + d = 1$. In case of M Level 1 classifiers, with $M > 2$, $\binom{M}{2} = \frac{M!}{(M-2)! \cdot 2!}$ relationship tables are produced. Afterwards, the diversity measure of a pair of classifiers is estimated by its corresponding 2×2 relationship table and the overall diversity of the group is the average value across all pairs.

	D_i correct (1)	D_i wrong (0)
D_i correct (1)	a	B
D_i wrong (0)	c	D

Table 27 2x2 relationship table of two classifiers

3.1.4.2.1.1 Q-statistic (Q)

Yule's Q statistic **[48]** for two classifiers , e.g., D_i and D_j, is

$$Q_{i,j} = \frac{a \cdot d - b \cdot c}{a \cdot d + b \cdot c} \tag{48}$$

For statistically independent classifiers, $Q_{i,j} = 0$. Q varies between -1 and 1 where the higher the value of the statistic, the lower the diversity between the classifiers. For a set of more than two classifiers the averaged Q statistic of all pairs is taken.

3.1.4.2.1.2 The Correlation Coefficient (p) [44]

The correlation between two binary classifier outputs with outputs as shown in Table 27 is:

$$p_{i,j} = \frac{a \cdot d - b \cdot c}{\sqrt{(a+b) \cdot (c+d) \cdot (a+c) \cdot (b+d)}} \tag{49}$$

When the classifiers are independent and uncorrelated, then $p_{i,j} = 0$. The higher the value of the Correlation coefficient, the lower diversity exists between the pair of classifiers.

3.1.4.2.1.3 Disagreement Measure (D) [44]

The disagreement measure is probably the most intuitive measure of diversity between a pair of classifiers. This measure is equal to the probability that the two classifiers will disagree on their decisions, that is:

$$D_{i,j} = b + c \tag{50}$$

Without calling it a disagreement measure, this statistic has been used in the literature for analysing classifier ensembles **[49]**, **[50]**. The higher the value of the disagreement measure, the lower the correlation between the two classifiers.

3.1.4.2.1.4 The Double-Fault Measure (DF)

The double fault measure **[51]** is another intuitive choice, as it gives the probability of classifiers D_i and D_j both being wrong,

$$DF_{i,j} = d \tag{51}$$

This measure is based on the concept that it is more important to know when simultaneous errors are committed than when both classifiers are correct. Thus the measure is related by design to the ensemble performance. The higher the value of this metric, the higher the correlation between the pair of classifiers.

3.1.4.2.1.5 Kappa Statistic (k)

Kappa Statistic **[52]** not only gives a measure of the degree of agreement, but it also has a test associated with it that can be employed to check if the apparent agreement cannot be attributed to change only. It is also helpful that the kappa statistic can show the level of agreement. Note that the proportion of agreement is given by $P_o = a + d$. The problem with this measure is that when two good classifiers are combined, they are bound to agree by correctly classifying most samples. Kappa statistic is given by:

$$k = \frac{2 \cdot (a \cdot d - b \cdot c)}{(a+b) \cdot (b+d) + (a+c) \cdot (c+d)} \tag{52}$$

A value of kappa below 0.40 is considered to represent poor agreement beyond chance, values between 0.40 and 0.75 indicate fair agreement, and values beyond 0.75 indicate excellent agreement.

3.1.4.2.2 Non-Pairwise Diversity Measures

The measures of diversity introduced below consider all the classifiers together and calculate directly one diversity value for the ensemble. For the non-pairwise measures we quote the formulae for L classifiers. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ be a labelled dataset, $x_j \in \Re^n$ coming from the classification problem in question. We can present the output of a classifier D_i an N-dimensional binary vector $\mathbf{y_i} = \left[y_{1,i}, y_{2,i}, \ldots, y_{N,i}\right]^T$, such that $y_{1,i} = 1$ if D_i recognises correctly x_j, and 0, otherwise, $i = 1, \ldots, L$.

3.1.4.2.2.1 The Entropy Measure (E) [44]

Intuitively, the ensemble is most diverse for a particular $x_j \in \mathbf{X}$ when $\lfloor L/2 \rfloor$ of the votes are zeros (or ones) and the other $L - \lfloor L/2 \rfloor$ votes are ones (or zeros). If they all were zeros or all were ones, there is no disagreement, and the classifiers cannot be deemed diverse. One possible measure of diversity based on this concept is:

$$E = \frac{1}{N} \cdot \frac{2}{L-1} \cdot \sum_{j=1}^{N} \min \left\{ \left(\sum_{i=1}^{L} y_{j,i} \right), \left(L - \sum_{i=1}^{L} y_{j,i} \right) \right\} \tag{53}$$

Entropy varies between 0 and 1, where 0 indicates no difference and 1 indicates the highest possible diversity.

3.1.4.2.2.2 Kohavi-Wolpert variance (kw)

The formulation of the proposed variance is taken from Kohavi's and Wolpert's paper **[53]**. They derived a decomposition formula for the error rate of a classifier, giving an expression of the variability of the predicted class label b for **x**, across training sets, within a specific classifier model as:

$$\mathrm{var}_x = \frac{1}{2} \left(1 - \sum_{i=1}^{c} P\left(b = \omega_i \mid x \right)^2 \right) \tag{54}$$

where $P\left(b = \omega_i \mid \mathbf{x}\right)$ is estimated as an average over different datasets. We use their general idea by looking at the variability of the predicted class label for **x** (for the given training set) using the classifier models $D_1, D_2, \ldots, D_L$. Instead of considering the class labels with the above technique, two possible classifier outputs are considered, correct and incorrect. $P\left(b = 1 \mid \mathbf{x}\right)$ and $P\left(b = 0 \mid \mathbf{x}\right)$ will be obtained as an average over D. If we denote by $l\left(\mathbf{x}_j\right)$ the number of classifiers from D that correctly recognize $\mathbf{x}_j$, i.e., $l\left(\mathbf{x}_j\right) = \sum_{i=1}^{L} y_{j,i}$ we obtain:

$$P\left(b = 1 \mid x\right) = \frac{l\left(x\right)}{L} \quad \textit{and} \quad P\left(b = 0 \mid x\right) = \frac{L - l\left(x\right)}{L} \tag{55}$$

Substituting (55) into (54),

$$\text{var}_x = \frac{1}{2}\left(1 - P\left(b = 1 \mid x\right)^2 - P\left(b = 0 \mid x\right)^2\right) \tag{56}$$

and averaging over the whole of the training set **X**, we obtain the kw measure of diversity as:

$$kw = \frac{1}{N \cdot L^2} \cdot \sum_{j=1}^{N} l\left(x_j\right) \cdot \left(L - l\left(x_j\right)\right) \tag{57}$$

3.1.4.3 Classifier Fusion Methods

Up to this point, we have a pool of Level 1 classifiers, and a criterion for the selection of a group of them that provides the highest diversity degree. Before we proceed with the classifier fusion approach, all the available combinations from the pool of classifiers are examined based on the correlation that they produce. Afterwards, the least-correlated group is selected for the implementation of the Level 2 classification.

Let $D = \left\{D_1, D_2, \ldots, D_L\right\}$ be the chosen set of classifiers and $\Omega = \left\{\omega_1, \ldots, \omega_c\right\}$ be a set of class labels. Despite the fact that our study focuses only on binary classification problems where $c = 2$, a more generalized form of the classifier ensembles will be given. Each classifier gets as its input a feature vector $x \in \Re^n$. The classifier output is a c-dimensional vector $D_i\left(x\right) = \left[d_{i,1}\left(x\right), \ldots, d_{i,c}\left(x\right)\right]^T$ where $d_{i,j}\left(x\right)$ is the degree of "support" given by classifier D_i to the hypothesis that **x** comes from class ω_j, $j = 1, \ldots, c$. Without loss of generality we can restrict $d_{i,j}\left(x\right)$ within the interval $\left[0,1\right]$, $i = 1, \ldots, L$, j=1,…,c, and call the classifier outputs "soft labels". Most often $d_{i,j}\left(x\right)$ is an estimate of the posterior probability $P\left(\omega_i \mid x\right)$.

Combining classifiers means we combine the L classifiers outputs $D_1\left(x\right), \ldots, D_L\left(x\right)$ to get a soft label for **x**, denoted $D\left(x\right) = \left[\mu_1\left(x\right), \ldots, \mu_c\left(x\right)\right]^T$. Table 28 gives our grouping of classifier fusion methods divided by the absence/presence of parameters to train at the fusion level and the type of classifier outputs.

First Level Outputs	Training at Fusion Level	
	No	Yes
Crisp	Majority Vote	Behavior-Knowledge Space "Naïve" Bayes
Soft	Minimum (Min) Maximum (Max) Mean Median Product	Probabilistic product Dempster-Shafer Decision Templates LDC classifier QDC classifier

Table 28 Classifier Fusion Techniques [44]

3.1.4.3.1 Decision Profiles

The degrees of support for a given input **x** can be interpreted in different ways, the two most common being confidences in the suggested labels and estimates of the posterior probabilities for the classes. Based on the above assumptions, the L classifier outputs for a particular input **x** can be organized in a decision profile ($DP(x)$) as the matrix **[44]**:

$$DP(\mathbf{x}) = \begin{bmatrix} d_{1,1}(\mathbf{x}) & \cdots & d_{1,j}(\mathbf{x}) & \cdots & d_{1,c}(\mathbf{x}) \\ d_{i,1}(\mathbf{x}) & \cdots & d_{i,j}(\mathbf{x}) & \cdots & d_{i,c}(\mathbf{x}) \\ d_{L,1}(\mathbf{x}) & \cdots & d_{L,j}(\mathbf{x}) & \cdots & d_{L,c}(\mathbf{x}) \end{bmatrix} \tag{58}$$

where the columns of the matrix represent the support from classifiers $D_1,\ldots,D_L$ for each class $\omega_1,\ldots,\omega_c$ and the rows, the output of each classifier $D_1(x),\ldots,D_L(x)$.

Some of the methods that will be described in the rest of this chapter use $DP(x)$ to find the overall support for each class and subsequently label the input **x** in the class with the largest support. There are two general approaches to this task. First, we can use the fact that the values in column j are the individual supports for class ω_j and derive an overall support value for that class. Denote by $\mu_j(\mathbf{x})$ the overall degree of support for ω_j given by the ensemble. Combination methods that use one column of $DP(x)$ at a time are called "class conscious" **[37]**. Examples from this group are the mean, min, max and the product. Alternatively, we may ignore the context of $DP(x)$ and treat the values

$d_{i,j}(\mathbf{x})$ as features in a new feature space, which we call the intermediate feature space. The final decision is made by another classifier that takes the intermediate feature space as input and outputs a class label. In **[37]** this class of methods is named "class-indifferent". Examples from this group are the Dempster-Shafer and the Decision Templates.

3.1.4.3.2 Crisp Labelling Fusion Methods

The below classifier ensemble techniques, require no further support from the Decision Profile of each input sample, and treat the predicted outcome of each classifier as a feature in a new dataset. Furthermore, the new feature dataset is based on the classification outcome, a crisp label that straightforward categorizes every sample to its predicted class. Highly used methods, belonging to this group, are outlined in the following chapters.

3.1.4.3.2.1 Majority Vote

Once the classifiers in the ensemble are trained, this combination method does not require any further training. For the majority vote combination, the class label assigned to **x** is the one that is most represented in the set of L crisp class labels obtained from $D_1(x),\ldots,D_L(x)$. Assume that $d_{i,j}(x)=1$ if D_i labels **x** in ω_j and 0 otherwise. The plurality vote will result in an ensemble decision for class ω_k if

$$\sum_{i=1}^{L} d_{i,k} = \max_{j=1}^{c} \sum_{i=1}^{L} d_{i,j} \tag{59}$$

Ties are resolved arbitrarily.

3.1.4.3.2.2 Naïve Bayes

Naïve Bayes **[54]** assumes that the classifiers are mutually independent and this is the reason the name "naïve" is used. For each classifier D_j, a $c \times c$ confusion matrix CM^j is calculated by applying D_j to the training data set. The $(k,s)th$ entry of this

matrix, $cm^j_{k,s}$ is the number of elements of the dataset whose true class label was k, and were assigned by D_j to class s.

By $cm^j_{\bullet,s}$ we denote the total number of elements labelled by D_j into class s (this is calculated as the sum of the s[th] column of CM^j). Using these values, a $c \times c$ label matrix LM^j is computed, whose $(k,s)th$ entry $lm^j_{k,s}$ is an estimate of the probability that the true label is k given that D_j assigns crisp class label s.

$$lm^j_{k,s} = \hat{P}\left(k \mid D_j(x) = s\right) = \frac{cm^j_{k,s}}{lm^j_{\bullet,s}} \tag{60}$$

For every $x \in \Re^n$, D_j yields a crisp label vector $D_j(x)$ pointing at one of the classes, say s in $\in \{1,\ldots,c\}$.

Associated with s is a soft label vector $\left[\hat{P}\left(1 \mid D_j(x) = s\right),\ldots,\hat{P}\left(c \mid D_j(x) = s\right)\right]^T$, which is the s[th] column of the label matrix LM^j. Let $s_1,\ldots,s_L$ be the crisp class labels assigned to **x** by classifiers $D_1,\ldots,D_L$, respectively. Then, by the independence assumption, the estimate of the probability that the true class label is i, (which is the i[th] component of the final label vector) is calculated by

$$\mu^i_{\hat{D}}(x) = \prod_{j=1}^{L} \hat{P}\left(i \mid D_j(x) = s_j\right) = \prod_{j=1}^{L} lm^j_{i,s_j},\ i = 1,\ldots,c \tag{61}$$

3.1.4.3.2.3 Behavior-Knowledge Space (BKS)

Behavior-Knowledge Space **[44]** is in fact a fancy name for the multinomial combination which requires training during the fusion level. Let $s = (s_1,\ldots,s_L) \in \Omega^L$ be the crisp class labels assigned to **x** by classifiers $D_1,\ldots,D_L$, respectively. Let also $Z = \{z_1,\ldots,z_N\}$, $z_j \in \Re^n$, be the crisply labelled training data set. Every possible combination of class labels $D_1(x)$, $(s_1,\ldots,s_L) \in \{1,\ldots,c\}^L$ is an index to a cell in a look-up table (the BKS table). The table is filled in using the data set $Z : z_j$ goes to the cell

indexed by $D_1(z_j),\ldots,D_L(z_j)$. Thus, each entry in the look-up table contains one of the following: a single class label (the one that is most often encountered amongst the elements of **Z** in this cell); no label (no element of **Z** had the respective combination of class labels); or a set of tied class labels (if more than one class have the same highest number of elements in this cell).

The decision for an $x \in \Re^n$ is made according to the class label of the cell indexed by $D_1(x),\ldots,D_L(x)$. Ties are broken randomly. If an empty cell is hit, the class label is chosen at random from $\{1,\ldots,c\}$. The operation of BKS is illustrated in Figure 18. BKS has sets of parameters that are estimated using the trained classifiers and the training data from the look-up table.

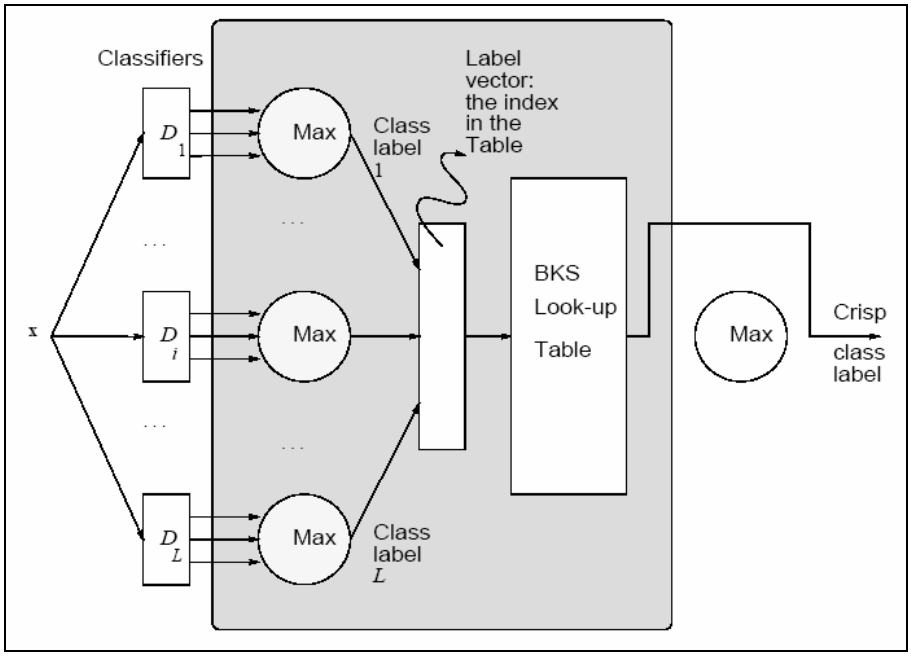

Figure 18 Behavior-Knowledge Space for Classifier Ensemble [44]

3.1.4.3.3 Soft Labelling Fusion Methods

Contrary to the Crisp labelling fusion techniques, the soft labelling methods require as an input, a decimal value usually in a probabilistic way of form. In other words, such techniques utilize the information that they get from the probabilistic

outcome of every classifier about the percentage that each sample belongs to its predicted class. Several soft labelling classifier ensembles are presented below.

3.1.4.3.3.1 Minimum, Maximum, Mean, Median, Product Rule

Once the classifiers in the ensemble are trained, these combination methods do not require any further training. Simple non-trainable combiners calculate the support for class ω_j using only the j[th] column of $DP(x)$ by

$$\mu_j(\mathbf{x}) = F\left[d_{1,j}(\mathbf{x}),\ldots,d_{L,j}(\mathbf{x})\right] \tag{62}$$

where F is a combination function. The class label of $\mathbf{x}$ is found as the index of the maximum $\mu_j(\mathbf{x})$. Minimum, Maximum, Mean, Median, Product Rule are given respectively by the following equations.

$$simple\ mean \rightarrow \mu_j(\mathbf{x}) = \frac{1}{L}\cdot\sum_{i=1}^{L} d_{i,j}(\mathbf{x}) \tag{63}$$

$$maximum\,/\,minimum\,/\,median \rightarrow \begin{cases} \mu_j(\mathbf{x}) = \max\limits_i \left\{d_{i,j}(\mathbf{x})\right\} \\ \mu_j(\mathbf{x}) = \min\limits_i \left\{d_{i,j}(\mathbf{x})\right\} \\ \mu_j(\mathbf{x}) = \operatorname{median}\limits_i \left\{d_{i,j}(\mathbf{x})\right\} \end{cases} \tag{64}$$

$$product \rightarrow \mu_j(\mathbf{x}) = \prod_{i=1}^{L} d_{i,j}(\mathbf{x}) \tag{65}$$

Figure 19 shows the operation of simple aggregation rules.

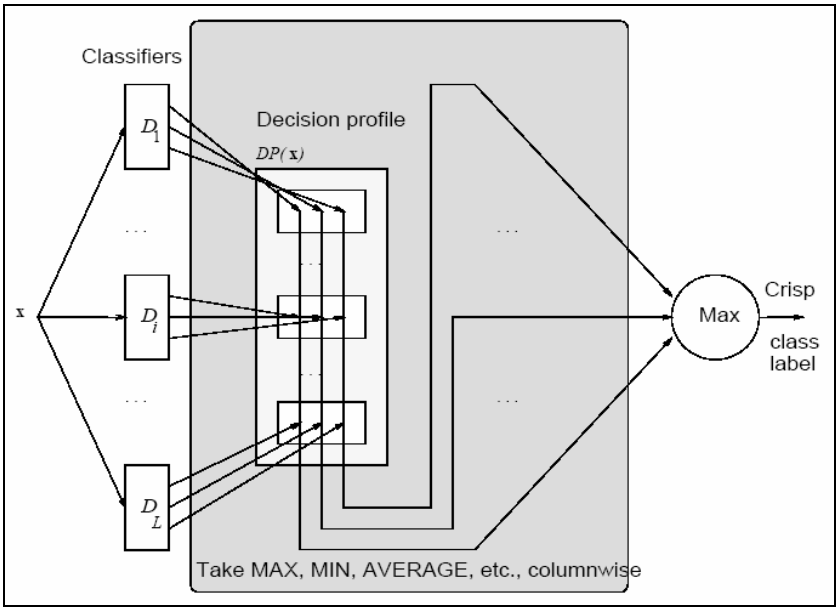

Figure 19 Simple Aggregation Rules for Classifier Ensemble [44]

3.1.4.3.3.2 Probabilistic Product

Probabilistic product **[55]** is an aggregation formula which gives the Bayes decision if the classifiers use mutually independent subsets of features and yield the true posterior probability, $d_{i,j}(x) = P(i \mid x_j)$, on their respective feature subspaces,

$$\mu_D^j(x) = \frac{\prod_{i=1}^{L} d_{i,j}(x)}{P(j)^{L-1}},\ j = 1,\ldots,c \tag{66}$$

For the prior probabilities $P(j)$ the sample based estimates from the training set **Z** were used.

$$\hat{P}(j) = \frac{N_j}{N},\ j = 1,\ldots,c \tag{67}$$

where N_j is the number of elements in **Z** from class j and N is the total training sample size. Even when the classifier outputs are not the true values but are estimates of the posterior probabilities, the probabilistic product works well as an aggregation connective.

3.1.4.3.3.3 Decision Templates [44]

The idea of the decision templates (DT) combiner is to create the most typical decision profile for each class ω_j, called the decision template, DT_j, and then compare it with the current decision profile $DP(\mathbf{x})$ using some similarity measure *S*. The closest match will label **x**. The decision Template DT_j for class ω_j is an $L \times c$ matrix with the average of the decision profiles of the elements of the training set **Z** labelled in class ω_j, and is given by the following equation:

$$DT_j = \frac{1}{N_j} \cdot \sum_{\substack{z_i \in \omega_j \\ z_i \in Z}} DP(z_i) \tag{68}$$

When $x \in \Re^n$ is submitted for classification, the Decision Template (DT) scheme matches $DP(\mathbf{x})$ to DT_i, $i = 1, \ldots, c$ and produces the soft class labels:

$$\mu_D^i(x) = S\left(DT_i, DP(x)\right), \; i = 1, \ldots, c \tag{69}$$

where S is interpreted as a similarity measure. The higher the similarity between the decision profile of the current **x** ($DP(\mathbf{x})$) and the decision template for class ω_j (DT_j), the higher the support for that class ($\mu_D^i(x)$). Two measures of similarity are based upon:

- The squared Euclidean distance $\left(DT(E)\right)$. The ensemble support for ω_j is

$$\mu_j(\mathbf{x}) = 1 - \frac{1}{L \times c} \cdot \sum_{i=1}^{L} \sum_{k=1}^{c} \left[DT_j(i,k) - d_{i,k}(\mathbf{x}) \right]^2 \tag{70}$$

  where $DT_j(i,k)$ is the $(i,k)th$ entry in decision template DT_j. Despite the squared Euclidean distance, any distance could be used, i.e., the Minkowski, Mahalanobis, and so on.

- A symmetric difference $\left(DT(S)\right)$. Symmetric difference comes from fuzzy set theory **[56]**, **[57]**. The support for ω_j is

$$\mu_j(\mathbf{x}) = 1 - \frac{1}{L \times c} \cdot \sum_{i=1}^{L} \sum_{k=1}^{c} \max\left\{ \min\left\{ DT_j(i,k), \left(1 - d_{i,k}(\mathbf{x})\right) \right\}, \min\left\{ \left(1 - DT_j(i,k)\right), d_{i,k}(\mathbf{x}) \right\} \right\} \tag{71}$$

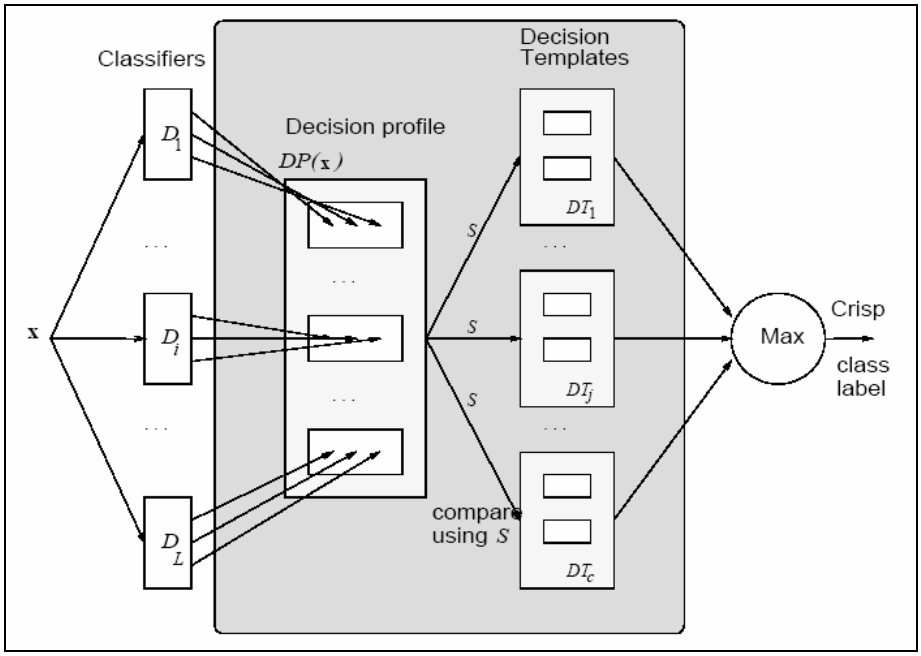

Figure 20 Decision Templates for Classifier Ensemble [44]

3.1.4.3.3.4 Dempster-Shafer Combination

This technique **[58]** is the one closest to the Decision Templates (DT). The classifier outputs $\{D_i(x)\}$ are probabilistic. Instead of calculating the similarity between the decision template DT_i and the decision profile $DP(x)$, the Dempster-Shafer algorithm goes further. The following steps are performed:

1. Let DT_j^i denote the i[th] row of the decision template for class ω_j. The "proximity" Φ between DT_j^i and $D_i(x)$ is calculated for every class $\omega_1,\ldots,\omega_c$ and for every classifier $i=1,\ldots,L$. As recommended in **[58]**, this proximity is calculated as

$$\Phi_{j,i}(x)=\frac{\left(1+\left\|DT_j^i-D_i(x)\right\|^2\right)^{-1}}{\sum_{k=1}^{c}\left(1+\left\|DT_k^i-D_i(x)\right\|^2\right)^{-1}} \tag{72}$$

   where $\|*\|$ is any matrix form.

2. Using (72), we calculate for every class $\omega_1,\ldots,\omega_c$ and for every classifier $i=1,\ldots,L$ the following belief degrees

$$b_j\left(D_i\left(x\right)\right)=\frac{\Phi_{j,i}\left(x\right)\cdot\prod_{k\neq j}\left(1-\Phi_{k,i}\left(x\right)\right)}{1-\Phi_{j,i}\left(x\right)\cdot\left[1-\prod_{k\neq j}\left(1-\Phi_{k,i}\left(x\right)\right)\right]} \quad (73)$$

3. The final DS label vector with membership degrees has the components

$$\mu_{\hat{D}}^{j}\left(x\right)=K\cdot\prod_{i=1}^{L}b_j\left(D_i\left(x\right)\right),\ j=1,\ldots,c \quad (74)$$

where K is a normalizing constant.

3.1.4.3.3.5 Classifiers acting as Classifiers ensemble

In the "class-indifferent" category some well known classifiers were also used: linear and quadratic discriminant classifiers (LDC and QDC **[59]**), and Fisher's discriminant (FSH) **[59]**. Generally, the fusion of the classifiers is a very flexible procedure in which every classifier can act like a classifier ensemble by taking as input the prediction of the Level 1 classifiers.

3.1.5 Feature Selection

Marker (feature) selection methods can be divided into two categories namely, filter and wrapper methods **[60]**. Filter methods focus on the intrinsic properties of data using various stochastic metrics such as Fisher's ratio (equation), T-statistics, information gain and many others. The attributes of the dataset are ranked according to how they score on such a measure and the highly ranked that give the highest classification accuracy are then selected as the most appropriate for further processing.

Wrapper methods on the other hand work in a recursive way, where a classifier is used to assign a relevance weight to each feature, and then the feature with the lowest weight is eliminated. A wrapper method as an iterative procedure was evaluated as a significant tool to assess the prediction accuracy of a classifier, in conjunction with the effectiveness of every feature that participates in the classification. The least significant attribute was eliminated after each cycle until only one feature was left. During the iterations weights are re-evaluated and potentially changing, while the process continues

recursively. At the end of the process, the smallest set of features achieving the highest classification accuracy is selected as the set of the revised dataset.

Note that filter methods could as well be applied recursively; however, feature weights stay stable from iteration to iteration. In these studies filter methods (mostly variations of Fisher's ratio) are used as the basic tool for marker selection, relying very much on the intrinsic characteristics of the selected features. The intrinsic data characteristic that domain experts are searching for through this metric is to identify differentially expressed genes in the two classes of interest. This aspect is not addressed by wrapper methods that, even with the exploitation of sophisticated pattern recognition tools, focus on classification accuracy neglecting intrinsic characteristics of the selected attributes.

4 Algorithmic Tools and Methodology

Our study focused on the classification analysis of the dataset based on both Support Vector Machines and Least Squares Support Vector Machines. Two powerful toolboxes written in Matlab Language **[61]**, the LS-SVMlab 1.5 toolbox **[62]**, and the OSUSVM toolbox **[63]**, respectively, supported our group with the implementation of the LS-SVMs and SVMs models.

The survival analysis of this study was carried out according to the SPSS **[64]** software. The Classifier Fusion techniques were implemented by one of our toolboxes, also written in Matlab, and published for public use in the Biopattern's web page **[65]**. The Hidden Space SVMs were also implemented by one of our toolboxes which is still under further development. We hope that the application of the HS-SVMs to a variety of biomedical datasets, under some modifications that we are currently working on in their structure, will be soon ready for publication. Finally, the feature selection method was also implemented by our group, and specifically from a PhD research which is also under development **[66]**.

A major fact in order to achieve a good classification performance is the ability of the classifier to absorb all the available information during its training. The performance of a classifier is a compound characteristic, whose most important component is the

classification accuracy. For that reason, widely used methods for non-straightforward prediction of the classification accuracy like the leave-one-out, the cross validation and the bootstrapping method **[67]** were available for the purposes of this study. The stratification of the samples was also followed by our study. Further explanation of these methods, and two other alternative cross validation techniques implemented by our group is also given below.

When the Long Term Analysis was under examination, we observed a difference between the number of samples categorized to class "**Complete Remission**" and these from class "**all others**". Specifically, 68.17% of the patients were categorized as those who achieved "**Complete Remission**", and 31.83% samples to class "**all others**" (see Table 29). This gap was much wider when the dataset was classified according to the Short Term Analysis (see Table 30).

Under this circumstance, in both analyses, the classifier was trained with much more information coming from the highly-numbered class than from the other. Having in mind that a good classifier generally requires a considerable amount of data from both classes in order to produce a good prognostic system, the partially lack of this information makes the system insufficient to classify unseen samples from the class that least informed the classifier. This was an open problem, frequently occurred in our previous studies, when several biomedical datasets with a significant difference between the populations of the two classes were classified. This issue seemed to affect the hyper-plane classifier to decide against the class with the least samples.

Taking into account the above problem, a possible solution was to select appropriate features with distinct characteristics that could separate the two classes. Unfortunately, such a solution was quite difficult, especially at the Short Term Analysis, because the classes didn't have a clear-cut meaning. In other words, class **"all others"** included outcomes "**Complete Remission**" and "**Resistant**", two different events with no big commonalities to their indicators, against class "**Induction death**" which on the contrary included patients with characteristics close to the characteristics of samples from class "**Resistant**". Finally, the solution to this problem was unexpectedly given by our alternative cross validation techniques which are analytically described below. Note that a possible solution was also provided by the implementation of **[68]**, in which a weighted

support vector machine sets a different penalty of misclassification for each training sample. Generally, despite the fact that the problem of uneven class sizes is very important, there is little consideration from the scientific community to the solution of this issue.

Complete Remission (CR)		
CR	**# of Patients**	**Percentage**
0	**162**	**31.83**
1	**347**	**68.17**

Table 29 Classes during Long Term Analysis

Induction Death		
inde	**# of Patients**	**Percentage**
1	**67**	**13.16**
0	**442**	**86.84**

Table 30 Classes during Short Term Analysis

4.1 Stratified K-fold Cross Validation

K-fold cross validation **[67]** is a well-known technique which was also adopted in our study for partitioning the dataset. In K-fold cross-validation, the original sample was randomly partitioned into K sub-samples. Of the K sub-samples, a single sub-sample was retained as the validation data for testing the model, and the remaining K−1 sub-samples were used as training data. The cross-validation process was then repeated K times (the number of the folds), with each of the K sub-samples used exactly once as the validation data. The K results from the folds then were averaged (or otherwise combined) to produce a single estimation.

Furthermore, the randomly partitioned folds were also stratified so that they could contain approximately the same proportions of labels as the original dataset. Afterwards,

the overall process was repeated many times and the final result was the average of all the K-fold Cross Validations. Despite the computationally expensive and the time-consuming procedure, the iteration of the entire procedure was essential in order to eliminate any bias that might have caused by the selection of the K-folds, and reliably estimate the accuracy of the classifier.

Note that the stratified K-fold cross validation gives no solution to the problem of the unbiased distribution between the two classes. The structure of the K-fold cross validation is straightforward, you can either choose to partition the entire set into K folds with stratification of the data or to randomly create the K folds no matter what is the label of the samples. In both cases, the number of samples from the entire dataset remains the same. Despite this fact, this technique was adopted to our study because is a widely used validation technique and a comparison between this method and our alternative cross validations will prove the optimization to the performance of the classifiers that is attained by our methods.

4.2 Stratified Cross Validation with permutation

Stratified Cross Validation with permutation was based on the simplest form of cross validation procedure **[67]**, training the classifier with a high percentage of the dataset and evaluates its performance with the rest of it. This method was experimentally proved as a non-reliable validation technique, because the classifier was trained and evaluated only one time, and the probability of choosing as an evaluation set a well-classified group of samples was really high. For that reason, this technique was enhanced by the stratification of the data, and was repeated many times. The extension to this method was based on the idea that the subset responsible for the training of the classifier could contain equivalent amount of data from both classes.

Generally, the cross-validation procedure requires a large subset, usually 80% of stratified data from both classes, for the training of the classifier and the remaining set for the evaluation of it. In our alternative method, the training set was chosen randomly from the entire dataset by selecting 80% from the least-populated class. The training set was also filled not with another 80% from the dense class, but with the same number of

samples that randomly selected from the first class. The remaining samples from the least-populated class as well as the samples from the other class constituted the evaluation set.

The drawback of this procedure is that a noticeable number of samples from the dense class didn't participate to the training of the classifier and the classifier might not get all the available information from it. The solution for this problem was given by the exhaustive iteration of the technique in order to eliminate the probability of having non-operating samples during the training procedure. A graphical representation is given in Figure 21.

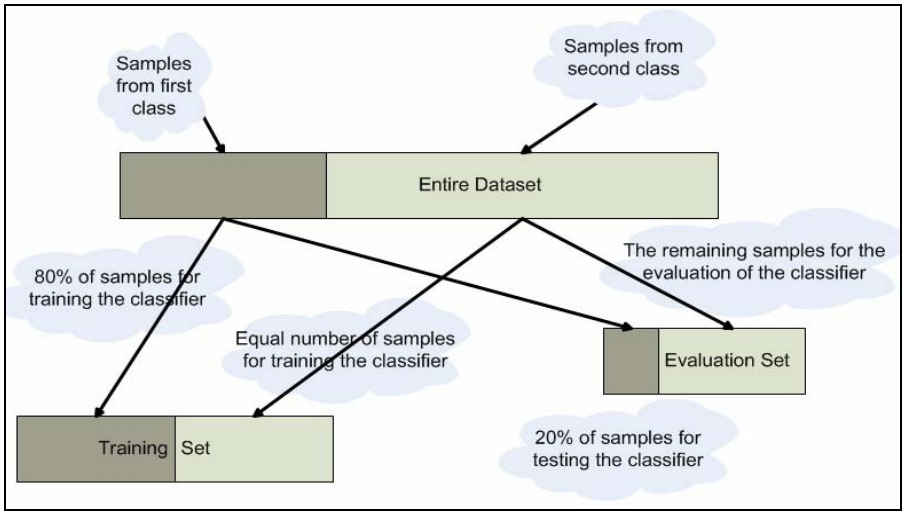

Figure 21 Stratified Cross Validation with permutation

4.3 Stratified K-fold Cross Validation with reproduction

Stratified K-fold Cross Validation with reproduction is an extensive form of the single Stratified K-fold Cross Validation **[67]**. The entire dataset was equally divided into K folds with stratification but the problem with the biased distribution of the samples remained.

After splitting the dataset into K-folds, our thought was to randomly reproduce the samples from the least-populated class in order to create an equivalent group of samples from both classes inside every fold. This technique was somehow based on the

existed “sampling with replacement” Cross Validation method **[67]**. Our group adopted this idea and implemented it to the Stratified K-Fold Cross Validation where the “sampling with replacement” acted individually inside each fold. Note that this technique like the previous one was also repeated many times, and the overall classification accuracy was the average value of the iterated estimated accuracies. The current technique is depicted below.

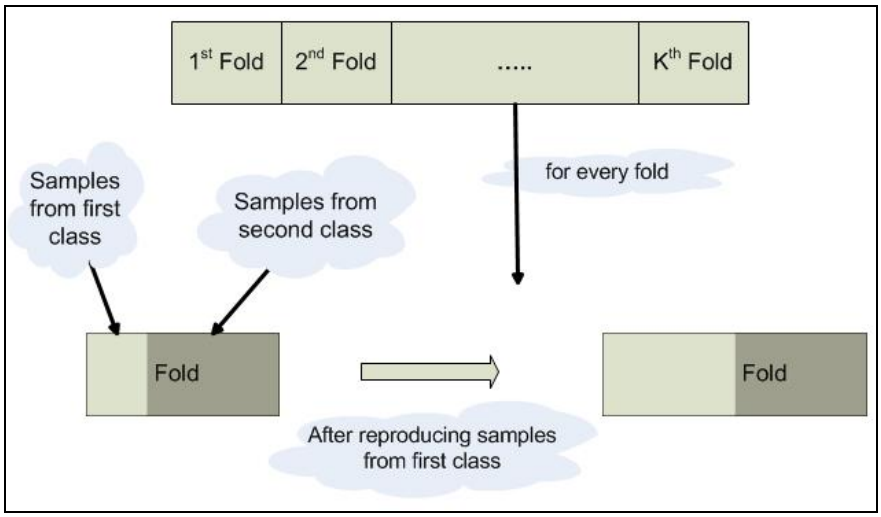

Figure 22 Stratified K-fold Cross Validation with reproduction

4.4 Classification Procedure

In this section, we represent the technical part of the classification, and the phases required for the training and the evaluation of the classifier’s prediction accuracy. The key for success in this process is to use as much as possible of the data to build the classifier (training), and also as much as possible unseen data to test its performance more thoroughly (independent test set). However, if we use all the data for training and the same data for testing, we might over-train the classifier so that it perfectly learns the available data and fails on unseen data. That is why it is important to have a separate data set on which to examine the final product.

For that reason, the entire dataset was separated into two subsets, 80% of the dataset was randomly splitted for the training part and the remaining set for testing the

overall classification approach. The smaller dataset, usually named as independent test set, was a subset that remained unseen during the training phase. Its scope was to assess the classifier's accuracy without causing any bias to the results.

On the other hand, the training set was responsible for adjusting the parameters of the classifiers that provide the best classification accuracy, and generally built a good prognostic system ready to classify the unseen dataset. In order to avoid any confusion with the terms "Training Set", "Evaluation Set" and "Test Set", it is essential to note that datasets "Training Set" and "Evaluation Set" through the training phase were completely different from datasets "Training Set" and "Test Set" at the testing part of the classification. The dataset participated to the training phase was actually the 80% of the initial dataset. This set was afterwards splitted, based on the cross validation techniques, into the "Training Set" and the "Evaluation Set", a sub-set which acted as pseudo-testing. Eventually, the classification procedure is outlined by the following Figure 23.

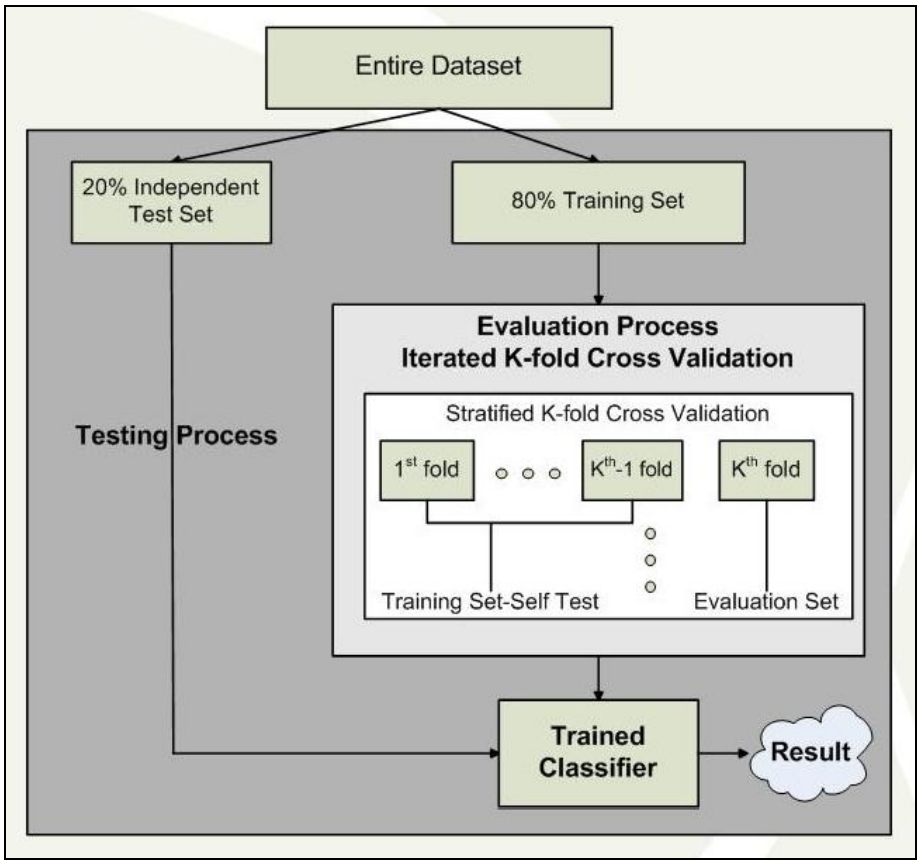

Figure 23 Graphical representation of the Classification procedure

Note that the procedure depicted above was exhaustively repeated and the final result of the classification was the average accuracy of all the iterations. Particularly, the initial dataset was randomly bootstrapped 20 times and as a result, twenty individual training and testing sets were generated for classification.

When the Least Square Support Vector Machines were implemented, the Radial Basis Function (RBF) Kernel was chosen as the Kernel function. On the other hand, the Support Vector Machines were implemented with a polynomial as a Kernel function. When the Hidden Space SVMs were applied, the Radial Basis Function, the Polynomial, and the Linear Kernel were chosen as a Hidden Kernel respectively. Our team has recently evaluated some alterations to the design and the implementation of the Hidden Kernels and we will be soon ready to publish the results of this research.

Finally, we would like to mention that the K-fold cross validation procedures were implemented with K equal to 5. A commonly used value for K is 10, but the morphology of the current dataset, didn't allow us to generate more folds.

4.5 Feature Selection Procedure

The current feature selection approach was a combination of the wrapper and the filter methods **[66]**. The feature selection problem is addressed by two types of approaches, the so called filter and wrapper methods. Wrapper methods are applied in an iterative fashion where the indicator's weights are re-evaluated and potentially changing from iteration to iteration. Filter methods on the other hand operate in a more static fashion; the weights remain stable and do not allow adjusting their performance in the next iteration cycle.

The filtering method, applied to this research, was based on the fisher ratio, a variation that has been widely applied by domain experts in leukemia problems **[70]**, and is given as:

$$f\left(g_i\right)=\frac{\left|\mu_+\left(g_i\right)-\mu_-\left(g_i\right)\right|}{\sigma_+\left(g_i\right)+\sigma_-\left(g_i\right)} \tag{75}$$

where $\mu_+\left(g_i\right),\mu_-\left(g_i\right)$ and $\sigma_+\left(g_i\right),\sigma_-\left(g_i\right)$ are the means and standard deviations of feature i in positive and negative classes respectively. Elaborating on the above equation

one could easily verify that features which differentiate more their expression in the two situations will be assigned higher weights than those that differentiate their expression less. On the bottom line, features with almost the same characteristics in the two situations will be assigned the minimum weight which is zero.

Such a filter criterion then focused on an intrinsic desirable characteristic of selected features from negative to positive class. We could apply Fisher's criterion in an iterative manner by ranking features according to how they score on it, eliminate low ranked features and re-evaluate scores for the survived features. However, in the next cycle, survived features could score the same scores as before since nothing has changed to the used parameters. That is, no matter how many iterations required, equation (75) always gave the same value for the survived features. Our goal then, was to apply such Fisher's metric in a wrapper manner where feature weights were potentially changing along the feature elimination process and thus improving performance over the original non wrapper filter method.

According to the SVMs theory, the direction vector $\mathbf{w}$ of the separating hyperplane, given by equation (16), was an expansion of those samples whose $\boldsymbol{\alpha}_i$ was non zero, i.e the support vectors. Based on this equation the individual components of the direction vector $\mathbf{w}$ could be found by:

$$w_i = \sum_{j=1}^{n} \alpha_j y_j x_{ij}, i = 1 \ldots m \tag{76}$$

Defining the contribution of each one of the n samples to $f(g_i)$ as a vector of n identical components we had:

$$\vec{f}(g_i) = \frac{1}{n}\left[\frac{\left|\mu_+(g_i) - \mu_-(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)}\right] \tag{77}$$

We then defined a new direction vector based on the support vectors (n_s) as follows:

$$w_i' = \sum_{j=1}^{n} sign(\alpha_j)\frac{1}{n_s}\left[\frac{\left|\mu_+(g_i) - \mu_-(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)}\right]_j = \frac{1}{n_s}\sum_{j=1}^{n} sign(\alpha_j)\left[\frac{\left|\mu_+(g_i) - \mu_-(g_i)\right|}{\sigma_+(g_i) + \sigma_-(g_i)}\right]_j \tag{78}$$

The direction vector $\boldsymbol{w}'$ whose components were defined by equation (78) defined a Fisher's hyperplane that passes through the origin and was derived using only the

support vector samples since their corresponding α values were non zero. This new hyperplane dynamically changed its direction vector across the feature selection process and hence, the weights of the associated features. Such a hyperplane could be used for defining a new ranking criterion which potentially changed its value along the process. Note that w' is an expansion of only those training samples whose α_i is non zero i.e. the support vectors. Also note that by using different kinds of kernels we were supplied with different sets of support vectors and thus, different Fisher lines.

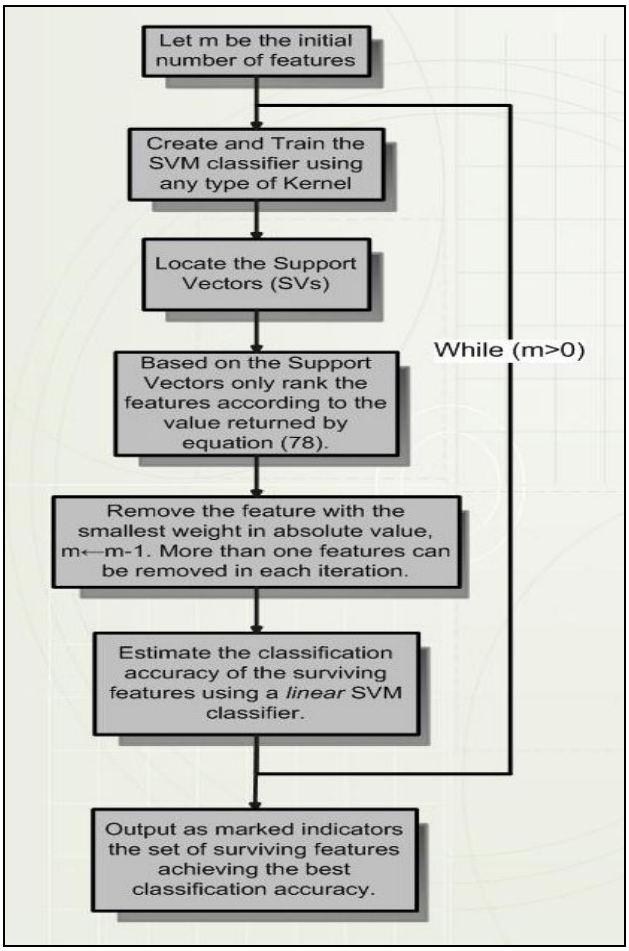

Figure 24 Feature Selection Algorithm

4.6 Classifiers Ensemble Procedure

The classifier fusion was a second-level of classification in which the classification results from the individually operated classifiers were organized in a new dataset for further classification (the decision profile of each sample). The Classifiers ensemble procedure was separated into two parts. In the first part, also named as Level 1 classification, a number of different classifiers (Level 1 classifiers) operated individually on the dataset. Afterwards, a group of them was selected based on several diversity measures, and their predicted outcome from Level 1 consisted the Decision Profile for each sample. Level 2 classification was the part in which several classifier fusion techniques operated on the label outputs of the Level 1 classifiers.

The Support Vector Machines (SVMs), the Least Squares SVMs and the Hidden Space SVMs, acted as Level 1 classifiers. Several highly-used classification algorithms were also implemented at Level 1, with the prospect to produce a large pool of diverse individual classifiers for fusion. In other words, we focused on the implementation of many different classifiers that provided acceptable classification results rather than having few classifiers that performed very well. PRTools **[69]**, a freeware pattern recognition toolbox, was also used for the needs of our study, and helped us utilize many of the Level 1 classifiers. Classifiers like the Probabilistic Neural Networks, the Fisher Discriminant Function, the Linear Dicriminant Function, the Naïve Bayes Classifier, the Quadratic Discriminant Function, the Radial Basis Neural Network Classifier, the K-Nearest Neighbor, and the Parzen Window Classifier contributed to the generation of the pool of Level 1 classifiers.

Every classifier that performed at Level 1 was evaluated not only for its prediction accuracy, but for its correlation degree with the other classifiers, as well. The diversity measures described above were all contributed to the assessment of the correlation degree of the entire ensemble. The average value of all the diversity measures was the criterion for the selection of the least correlated group of classifiers. Note that all the possible combinations of the available classifiers were examined and their diversity degree is outlined below. The entire procedure is depicted in Figure 25.

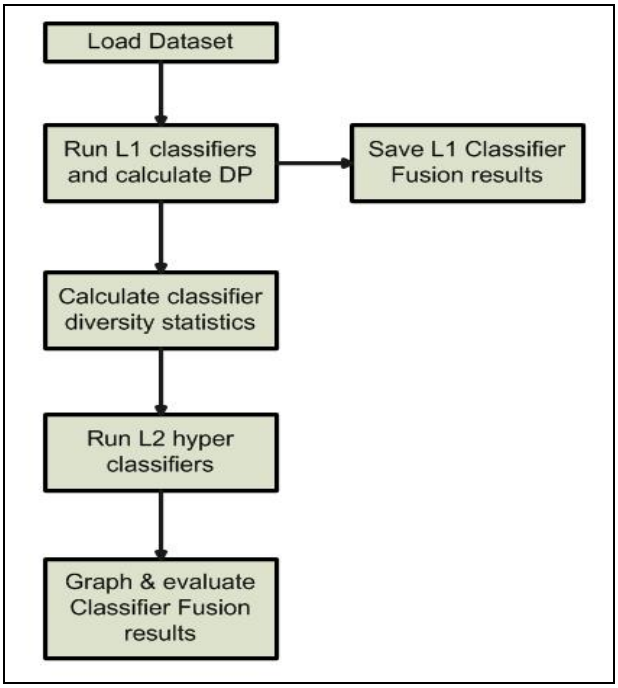

Figure 25 Classifiers Ensemble Procedure

5 Results

Below we represent the classification as well as the survival analysis results from our research. In the following chapters we are going to give an analytical explanation of the measures that we used to assess the classification accuracy, the formulation of the tables that present the results, and a thorough discussion of the results.

5.1 Supervised Learning Results

5.1.1 Presentation of the results

The results from the classification approach are categorized into Short Term and Long Term Analysis results. As shown in Figure 23, the classification procedure was splitted into two parts, the training and the testing part. The results from both phases are given by the following tables.

The classification results are outlined separately into two different tables. In the first table, the classification results from the training phase are given, whereas in the second table we represent the accuracy of each classifier when operated on the test set. Both tables are additionally separated into three parts. In the first part, the best set of the classification parameters is given. In case of Least Square Support Vector Machines, **Gam.** and **Sig.** are respectively the regularization parameter, determining the trade-off between the fitting error minimization and smoothness of the estimated function, and the parameter of the RBF Kernel.

In case of Support Vector Machines, **Deg.** and **Gam.** are respectively the Degree of the polynomial Kernel and the regularization parameter of the classifier. In the middle part and the third part, we represent the results of the training and the testing process. To find out how the errors are distributed along the classes we constructed a confusion matrix with measures **TP**, **FP**, **TN**, **FN**, which are the True Positives, False Positives, True Negatives and False Negatives predictions of the classifier. The **ROC** value represents the Area Under the ROC curve. Specifically, four different measures, based on the **TP**, **FP**, **TN**, **FN**, measures were depicted. These were the Specificity, Sensitivity, the False Positive Rate, and the False negative Rate, given by equation (79).

Sensitivity is the proportion of people that tested positive of all the positive people tested; that is (true positives) / (true positives + false negatives). It can be seen as the probability that the test is positive given that the patient is sick. The higher the sensitivity, the fewer real cases of diseases go.

Specificity is the proportion of people that tested negative of all the negative people tested; that is (true negatives) / (true negatives + false positives). As with sensitivity, it can be looked at as the probability that the test is negative given that the patient is not sick. The higher the specificity, the fewer healthy people are labeled as sick.

$$\begin{cases} Sensitivity\ or\ True\ Positive\ Rate\ (TPR) = \dfrac{TP}{TP + FN} \\ False\ Positive\ Rate\ (FPR) = \dfrac{FP}{FP + TN} \\ Specificity\ or\ True\ Negative\ Rate\ (TNR) = \dfrac{TN}{TN + FP} \\ False\ Negative\ Rate\ (FNR) = \dfrac{FN}{FN + TP} \end{cases} \tag{79}$$

The top part of the table informs the reader about the tested dataset and the Cross Validation procedure that was applied to it during the training phase. The positive class when the Short Term analysis was under examination was the "**all others**" class and the negative class was the "**Death**" event. The positive class in case of Long Term Analysis corresponded to the achievement of Complete Remission "**CR**", whereas the negative class was class "**all others**". Lastly we would like to mention once again that all the results were the averaged iterated classification procedure performances.

We would also like to inform the reader that we didn't analytically present the classifiers performance provided by the Prtools toolbox, because their only scope was to contribute to the production of a large pool of classifiers for the classifiers ensemble. At the bottom of this chapter, there is a total representation of the performance of each classifier, and their confidence interval.

5.1.2 Short Term Analysis Results

Due to the fact that the overall classification approach was very exhaustive, with five different datasets participating to the classification, Short Term and Long term analysis under examination, and the iteration process through the training and the testing procedure, the representation of every classification result was very difficult. For that reason, this study outlines the most accurate results provided by the several classification techniques.

5.1.2.1 Short Term Analysis results with LS-SVM

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	5	100,00%	82,29%	17,71%	0,00%	99,80%	98,11%	1,89%	0,20%	50,84%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	5	100,00%	85,37%	14,63%	0,00%	99,64%	96,67%	3,33%	0,36%	51,48%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	7	97,53%	24,92%	75,08%	2,47%	82,13%	70,62%	29,38%	17,87%	55,76%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	7	97,30%	27,54%	72,46%	2,70%	82,28%	71,17%	28,83%	17,72%	55,56%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	72,07%	32,11%	67,89%	27,93%	70,65%	49,31%	50,69%	29,35%	60,71%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	71,76%	32,44%	67,56%	28,24%	68,97%	45,19%	54,81%	31,03%	61,89%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	10	100,00%	86,52%	13,48%	0,00%	99,71%	99,54%	0,46%	0,29%	50,09%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	10	100,00%	88,80%	11,20%	0,00%	99,31%	99,00%	1,00%	0,69%	50,16%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	95,25%	56,70%	43,30%	4,75%	87,46%	74,56%	25,44%	12,54%	56,45%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	94,21%	54,66%	45,34%	5,79%	86,68%	72,33%	27,67%	13,32%	57,17%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	35	72,11%	31,48%	68,52%	27,89%	70,57%	46,68%	53,32%	29,43%	61,94%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	35	72,04%	28,63%	71,37%	27,96%	70,62%	49,50%	50,50%	29,38%	60,56%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	15	99,96%	87,13%	12,87%	0,04%	99,35%	98,98%	1,02%	0,65%	50,19%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	15	99,98%	88,26%	11,74%	0,02%	99,23%	98,33%	1,67%	0,77%	50,45%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	95,62%	49,44%	50,56%	4,38%	87,14%	70,03%	29,97%	12,86%	58,56%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	93,85%	47,01%	52,99%	6,15%	83,84%	69,06%	30,94%	16,16%	57,39%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	76,07%	27,85%	72,15%	23,93%	73,61%	44,96%	55,04%	26,39%	64,33%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	75,89%	29,61%	70,39%	24,11%	72,50%	42,33%	57,67%	27,50%	65,08%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	7	99,99%	62,92%	37,08%	0,01%	98,84%	96,98%	3,02%	1,16%	50,93%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	7	100,00%	65,22%	34,78%	0,00%	98,72%	98,00%	2,00%	1,28%	50,36%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	96,11%	47,23%	52,77%	3,89%	87,57%	70,70%	29,30%	12,43%	58,44%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	95,08%	48,91%	51,09%	4,92%	85,11%	67,15%	32,85%	14,89%	58,98%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	35	76,24%	28,67%	71,33%	23,76%	74,26%	43,67%	56,33%	25,74%	65,34%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	35	76,58%	4,79%	95,21%	23,42%	74,67%	37,00%	63,00%	25,33%	68,83%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	npm	d835

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	20	99,99%	86,29%	13,71%	0,01%	99,47%	99,24%	0,76%	0,53%	50,12%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	20	100,00%	90,88%	9,12%	0,00%	99,83%	96,25%	3,75%	0,17%	51,79%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	25	96,88%	34,74%	65,26%	3,12%	82,79%	70,52%	29,48%	17,21%	56,13%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	25	96,53%	39,46%	60,54%	3,48%	84,93%	68,15%	31,85%	15,08%	58,39%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	76,07%	22,92%	77,08%	23,93%	73,18%	49,15%	50,85%	26,82%	62,05%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	75,37%	24,39%	75,61%	24,63%	76,38%	43,38%	56,63%	23,62%	66,50%

Up to this point, a brief discussion about the performance of the LS-SVM is essential. We observed that all the available datasets from the AML99 protocol appeared to have corresponded really well to the implementation of our alternative validation approaches. The cross validation techniques with permutation and reproduction improved the degree of the area under the ROC curve, and succeeded to raise the Specificity value. Specifically, the cross validation with reproduction provided the best classification accuracy in all cases.

5.1.2.2 Short Term Analysis results with SVMs

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	1	100,00%	0,00%	100,00%	0,00%	88,41%	75,44%	24,56%	11,59%	56,52%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	1	100,00%	0,00%	100,00%	0,00%	88,36%	71,67%	28,33%	11,64%	58,35%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	7	100,00%	0,00%	100,00%	0,00%	69,24%	54,42%	45,58%	30,76%	57,41%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	7	100,00%	0,00%	100,00%	0,00%	68,02%	51,77%	48,23%	31,98%	58,13%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	1	100,00%	0,00%	100,00%	0,00%	88,42%	75,96%	24,04%	11,58%	56,24%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	1	100,00%	0,00%	100,00%	0,00%	88,36%	71,67%	28,33%	11,64%	58,35%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
4	20	100,00%	0,00%	100,00%	0,00%	88,82%	80,96%	19,04%	11,18%	53,93%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
4	20	100,00%	0,00%	100,00%	0,00%	88,24%	84,00%	16,00%	11,76%	52,12%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	30	100,00%	0,00%	100,00%	0,00%	67,84%	61,50%	38,50%	32,16%	53,17%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	30	100,00%	0,02%	99,98%	0,00%	66,31%	61,34%	38,66%	33,69%	52,49%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	20	100,00%	0,00%	100,00%	0,00%	89,30%	81,48%	18,52%	10,70%	53,91%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	20	100,00%	0,00%	100,00%	0,00%	89,31%	80,00%	20,00%	10,69%	54,66%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	7	100,00%	0,00%	100,00%	0,00%	87,84%	86,22%	13,78%	12,16%	50,80%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	7	100,00%	0,00%	100,00%	0,00%	88,08%	84,17%	15,83%	11,92%	51,96%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	10	100,00%	0,00%	100,00%	0,00%	68,51%	66,90%	33,10%	31,49%	50,81%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	10	100,00%	0,00%	100,00%	0,00%	66,52%	56,72%	43,28%	33,48%	54,90%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	1	100,00%	0,00%	100,00%	0,00%	87,81%	85,91%	14,09%	12,19%	50,97%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	1	100,00%	0,00%	100,00%	0,00%	88,08%	84,17%	15,83%	11,92%	51,96%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	2	100,00%	0,00%	100,00%	0,00%	87,08%	83,76%	16,24%	12,92%	51,70%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	2	100,00%	0,00%	100,00%	0,00%	87,31%	76,00%	24,00%	12,69%	55,65%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	10	100,00%	0,00%	100,00%	0,00%	67,22%	64,17%	35,83%	32,78%	51,53%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	10	100,00%	0,00%	100,00%	0,00%	64,28%	58,45%	41,55%	35,72%	52,92%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
4	20	100,00%	0,00%	100,00%	0,00%	85,57%	82,37%	17,63%	14,43%	51,62%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
4	20	100,00%	0,00%	100,00%	0,00%	84,87%	72,00%	28,00%	15,13%	56,44%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	npm	d835

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	25	100,00%	0,00%	100,00%	0,00%	83,38%	84,71%	15,29%	16,62%	49,46%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	25	100,00%	0,00%	100,00%	0,00%	83,17%	81,25%	18,75%	16,83%	50,96%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	2	100,00%	0,00%	100,00%	0,00%	62,86%	64,88%	35,12%	37,14%	48,99%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	2	100,00%	0,00%	100,00%	0,00%	61,77%	62,55%	37,45%	38,23%	49,61%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	20	100,00%	0,00%	100,00%	0,00%	82,99%	83,79%	16,21%	17,01%	49,56%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	20	100,00%	0,00%	100,00%	0,00%	83,17%	81,25%	18,75%	16,83%	50,96%

On the contrary, the Support Vector Machines didn't correspond well to all the cross validation techniques. The only improvement that was achieved was to balance the misclassification between the uneven class sizes by increasing the specificity degree and to keep at the same time the area under the ROC curve at the same levels.

5.1.2.3 Short Term Analysis results with HS-SVMs

- ***Variables: All***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol

Stratified K-Fold Cross Validation---All Combinations										
TRAINING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Evaluation Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%
TESTING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Independent Test Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on

Stratified K-Fold Cross Validation---All Combinations										
TRAINING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Evaluation Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%
TESTING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Independent Test Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd

Stratified K-Fold Cross Validation---All Combinations										
TRAINING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Evaluation Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%
TESTING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Independent Test Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm	itd	d835

Stratified K-Fold Cross Validation---All Combinations											
TRAINING PHASE											
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Evaluation Set					
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC	
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%	
TESTING PHASE											
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Independent Test Set					
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC	
-	-	100,00%	0,00%	100,00%	0,00%	100,00%	100,00%	0,00%	0,00%	50,00%	

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	npm	d835

Stratified K-Fold Cross Validation---Kernel RBF---Hidden Space Kernel RBF										
TRAINING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Evaluation Set				
Sigma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	0,1	100,00%	0,00%	100,00%	0,00%	100,00%	90,00%	10,00%	0,00%	55,00%
TESTING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Independent Test Set				
Sigma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	0,1	100,00%	0,00%	100,00%	0,00%	100,00%	92,50%	7,50%	0,00%	53,75%

The Hidden Space SVMs were totally misclassified all the datasets when the Short term Analysis was under examination. Several Hidden kernels were applied to the datasets but it wasn't possible to obtain a good separation between the classes.

5.1.2.4 Short Term Analysis results with Wrapper Method

The feature selection results are given by the following tables. In the first table, the rows represents each iteration from the wrapper approach; starting from the entire dataset to the most significant feature that remained at the end of the process. The row with the number of features that provided the best results is highlighted. Several measures like the Area Under the ROC curve (AUROC), the Specificity and Sensitivity, the Success Rate which shows the number of samples that correctly classified, the Self

Accuracy which provides how well the classifier was trained, and the number of Support Vectors, informed the reader about the classification performance of each group.

# of Features	Success Rate	AUROC	Self Accuracy	Sensitivity	Specificity	SVs (-1)	SVs (1)
10	79,66%	55,40%	100,00%	86,79%	24,00%	21	50
9	78,41%	52,08%	100,00%	86,15%	18,00%	21	42
8	77,27%	51,44%	99,98%	84,87%	18,00%	21	38
7	77,95%	52,69%	99,67%	85,38%	20,00%	20	37
6	81,25%	56,29%	97,98%	88,59%	24,00%	20	37
5	83,30%	55,71%	95,60%	91,41%	20,00%	21	39
4	85,80%	53,19%	93,19%	95,38%	11,00%	21	36
3	86,25%	50,83%	91,49%	96,67%	5,00%	22	40
2	87,50%	50,23%	89,95%	98,46%	2,00%	23	55
1	88,41%	49,87%	89,33%	99,74%	0,00%	23	53

The feature selection technique was repeated many times, according to the bootstrapping separation of the dataset. The Training and the Testing dataset was randomly generated from the entire dataset, and the current approach was tested many times (20 iterations). To get the final classification accuracy we averaged the estimates of all the iterations. Every time that the wrapper technique was implemented, the approach provided information about the least significant feature which was eliminated after each step. At the end of the process, the dataset was only consisted of the most significant feature and the occurred frequency of all the features had the formulation of the following table:

	Occurred Frequency:						
Features	**1**	**2**	**3**	**4**	**Most Significant**	**….**	**m**
Start of Process	100%	100%	100%	100%	**100%**	100%	100%
⋮							
End of Process	0%	0%	0%	0%	**100%**	0%	0%

After the repetition of the feature selection, all the available information was fused to the table below. This table provides significant information about the frequency of every feature that participated in each step of the wrapper technique, after the twenty

iterations of the process. We observed that when the number of features was reduced to six, the highest area under the ROC curve was achieved. At this step, the features with the higher weight value were:

wbc_dia	PS_dia	citomol	PS_dia	Sex	Hb_On

After 20 Iterations		Number of Features									
		1	2	3	4	5	6	7	8	9	10
Frequency of Every Feature	Sex	15%	15%	25%	40%	50%	60%	85%	95%	100%	100%
	wbc_dia	30%	45%	60%	70%	85%	85%	90%	95%	95%	100%
	PS_dia	5%	25%	45%	65%	70%	85%	90%	95%	100%	100%
	bl_bm_dia	5%	15%	15%	25%	35%	50%	75%	90%	100%	100%
	Hb_On	5%	10%	15%	20%	50%	60%	70%	80%	85%	100%
	PLTS_On	30%	30%	55%	65%	70%	80%	85%	90%	90%	100%
	Citomol	10%	40%	65%	70%	80%	85%	90%	100%	100%	100%
	exm_on	0%	0%	0%	0%	5%	10%	20%	40%	55%	100%
	Itd	0%	15%	15%	20%	25%	45%	45%	60%	100%	100%
	d835	0%	5%	5%	25%	30%	40%	50%	55%	75%	100%

5.1.2.5 Short Term Analysis results with Classifier Fusion

The classifier fusion results are presented by the following tables. The correlation degree of all the possible combinations of the L1 classifiers is assessed based on the average value of the diversity measures, described in chapter 3.1.4.2. Due to the high number of cases, the tables below represent only the classification results by all the available Level 2 classifiers when the least correlated group of Level 1 classifiers participated to the fusion procedure. The classification performance of the above Level 2 classifiers was also compared with the performance of the fusion techniques when all the Level 1 classifiers contributed to the classifier ensemble. Moreover, in this study we examined the case in which the correlation metrics were not taken into consideration. In other words, the fusion classifiers were applied to all the available combinations of the Level 1 classifiers and the best accuracies by all the Level 2 classifiers are also depicted below. Note that the Classifier fusion technique was implemented in part by a graduate student of our group.

Classifier Fusion based on the best AUROC	Level 2 Classifier	Combination of L1 Cls	Sensitivity	FPR	Specificity	FNR	AUROC	Diversity
	Majority Vote	LDC--NBC--LSSVM	76,67%	40,00%	60,00%	23,33%	68,34%	0,459
	Minimum	LDC--LSSVM	72,69%	38,00%	62,00%	27,31%	67,35%	0,263
	Maximum	LDC--LSSVM	72,69%	38,00%	62,00%	27,31%	67,35%	0,263
	Average	LDC--LSSVM	72,69%	38,00%	62,00%	27,31%	67,35%	0,263
	Product Rule	LDC--LSSVM	72,69%	38,00%	62,00%	27,31%	67,35%	0,263
	Decision Templates	LDC--RBNC--LSSVM	74,62%	40,00%	60,00%	25,38%	67,31%	0,624
	Fisher	LDC--QDC--NBC--Fisher	95,89%	86,00%	14,00%	4,11%	54,95%	0,536
	KNN	NBC--RNNC--LSSVM	95,77%	82,00%	18,00%	4,23%	56,89%	0,576
	LDC	NBC--RNNC	92,18%	66,00%	34,00%	7,82%	63,09%	0,418
	QDC	KNNC--NBC	80,26%	69,00%	31,00%	19,74%	55,63%	0,404
	RBNC	LDC--LSSVM	94,62%	85,00%	15,00%	5,38%	54,81%	0,263
	RNNC	NBC--Fisher	96,03%	81,00%	19,00%	3,97%	57,52%	0,391

Classifier Fusion based on the best averaged Diversity	Level 2 Classifier	Combination of L1 Cls	Sensitivity	FPR	Specificity	FNR	AUROC	Diversity
	Majority Vote	LDC--RBNC	85,90%	65,00%	35,00%	14,10%	60,45%	0,822
	Minimum	LDC--RBNC	78,08%	49,00%	51,00%	21,92%	64,54%	0,822
	Maximum	LDC--RBNC	78,08%	49,00%	51,00%	21,92%	64,54%	0,822
	Average	LDC--RBNC	78,08%	49,00%	51,00%	21,92%	64,54%	0,822
	Product Rule	LDC--RBNC	78,08%	49,00%	51,00%	21,92%	64,54%	0,822
	Decision Templates	LDC--RBNC	74,36%	41,00%	59,00%	25,64%	66,68%	0,822
	Fisher	LDC--RBNC	99,49%	100,00%	0,00%	0,51%	49,75%	0,822
	KNN	LDC--RBNC	99,62%	100,00%	0,00%	0,38%	49,81%	0,822
	LDC	LDC--RBNC	89,62%	90,00%	10,00%	10,38%	49,81%	0,822
	QDC	LDC--RBNC	84,62%	85,00%	15,00%	15,38%	49,81%	0,822
	RBNC	LDC--RBNC	99,74%	100,00%	0,00%	0,26%	49,87%	0,822
	RNNC	LDC--RBNC	99,62%	100,00%	0,00%	0,38%	49,81%	0,822

Classifier Fusion with all the available L1 classifiers	Level 2 Classifier	Combination of L1 Cls	Sensitivity	FPR	Specificity	FNR	AUROC	Diversity
	Majority Vote	all L1 classifiers	95,51%	88,00%	12,00%	4,49%	53,76%	0,503
	Minimum	all L1 classifiers	90,13%	78,00%	22,00%	9,87%	56,07%	0,503
	Maximum	all L1 classifiers	89,87%	78,00%	22,00%	10,13%	55,94%	0,503
	Average	all L1 classifiers	92,69%	78,00%	22,00%	7,31%	57,35%	0,503
	Product Rule	all L1 classifiers	92,05%	77,00%	23,00%	7,95%	57,53%	0,503
	Decision Templates	all L1 classifiers	80,39%	66,00%	34,00%	19,61%	57,20%	0,503
	Fisher	all L1 classifiers	99,49%	100,00%	0,00%	0,51%	49,75%	0,503
	KNN	all L1 classifiers	95,00%	90,00%	10,00%	5,00%	52,50%	0,503
	LDC	all L1 classifiers	60,13%	60,00%	40,00%	39,87%	50,07%	0,503
	QDC	all L1 classifiers	55,13%	55,00%	45,00%	44,87%	50,07%	0,503
	RBNC	all L1 classifiers	100,00%	100,00%	0,00%	0,00%	50,00%	0,503
	RNNC	all L1 classifiers	97,56%	99,00%	1,00%	2,44%	49,28%	0,503

5.1.2.6 Discussion to the Short Term Analysis Results

After the implementation of many classification techniques, several observations are enjoying our attention. First of all, we must mention that the Short Term Analysis was experimentally proved as a very difficult classification problem. Our assumptions about the difficulties that we were going to face up were verified by the classification results. The wide gap between the uneven class sizes, finally affected the classifiers to decide against the least numbered class.

Highly used cross validation techniques provided non-relevant support for a good separation of the samples. Despite the fact that the Sensitivity measure was fairly good all the time, the percentage of the Specificity from the majority of the classifiers was below 50% and sometimes equal to zero. The Least Squares SVMs and especially the Hidden Space SVMs were proved inefficient to deal with this classification. Furthermore, even if the classifier ensemble techniques are generally characterized by their ability to improve the performance of the classification, this time the highly correlated Level 1 classifiers provided no relevant information to the classifier ensembles. No good results were also given by the Wrapper technique. The selection of the most significant group of features through the combination of the Fisher Ratio and the classification accuracy didn't succeed to improve the classification.

Summing up, the best classification result was provided by the Support Vector Machines with polynomial Kernel of degree 6, when the classifier was trained with the K-fold Cross Validation.

Stratified K-Fold Cross Validation---Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
88,36%	71,67%	28,33%	11,64%	58,35%

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
99,64%	96,67%	3,33%	0,36%	51,48%

When the SVMs and LS-SVMs were trained based on our alternative cross validation techniques we observed an improvement to the performance of these classifiers. Precisely, when the iterated stratified cross validation with permutation was

applied to the SVM classifier the area under the ROC curve remained at the same levels, but there was also a high increase to the Specificity measure. In other words, the predicted system was now capable to separate at least half of the samples that belong to the least-numbered class. The only negative reaction was the reduction of the Sensitivity value.

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
68,02%	51,77%	48,23%	31,98%	58,13%

Hopefully, the improvement of the LS-SVM classifier's performance when the model was trained by the iterated stratified K-fold cross validation with reproduction was above all expectations. The LS-SVM provided the best classification accuracy above all, and the area under the ROC curve reached the 70%. The reproduction of the data helped the Radial Basis Function (RBF) Kernel to separate a significant number of unseen samples from both classes, and balanced the accuracy of the Sensitivity and specificity measure. Below we represent the best classification result for the Short Term Analysis that was provided by features:

Sex	**wbc_dia**	**PS_dia**	**bl_bm_dia**	**Hb_On**	**PLTS_On**	**citomol**	**exm_on**	**itd**	**d835**

and a comparison between the K-fold cross validation and our alternative technique with the reproduction.

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
98,72%	98,00%	2,00%	1,28%	50,36%
Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
74,67%	37,00%	63,00%	25,33%	68,83%

Another interesting issue was the number of the Support Vectors that required for the SVMs classifier in order to separate the classes when a Kernel or a Hidden Space Kernel function was implemented. Zhang **[36]**, asserts that in many cases the Hidden

Space SVM require less Support Vectors than a simple SVM model to classify, with almost the same performance, a dataset. This reduction to the Support Vectors often eliminate the risk of getting overfitting and such event is very important in the classification analysis.

Unfortunately, the HS-SVM classifiers, at this part of the research, provided no good performance and a comparison between the performance of the HS-SVM and the simple case of the SVMs had no meaning.

5.1.3 Long Term Analysis results

5.1.3.1 Long Term Analysis results with LS-SVMs

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	94,93%	58,91%	41,09%	5,07%	91,85%	64,95%	35,05%	8,15%	63,45%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	95,22%	59,00%	41,00%	4,78%	93,29%	65,00%	35,00%	6,71%	64,15%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	93,87%	57,10%	42,90%	6,13%	90,73%	62,46%	37,54%	9,27%	64,14%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	94,06%	57,30%	42,70%	5,94%	92,17%	64,13%	35,87%	7,83%	64,02%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,5	30	74,14%	32,49%	67,51%	25,86%	71,24%	39,25%	60,75%	28,76%	65,99%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,5	30	73,25%	33,47%	66,53%	26,75%	73,84%	39,92%	60,08%	26,16%	66,96%

- ***Variables:***

Sex	Wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	95,30%	59,04%	40,96%	4,70%	91,63%	67,30%	32,70%	8,37%	62,16%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	30	95,24%	60,00%	40,00%	4,76%	93,08%	63,61%	36,39%	6,92%	64,73%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	93,87%	56,19%	43,81%	6,13%	90,24%	63,91%	36,09%	9,76%	63,16%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	93,87%	56,63%	43,37%	6,13%	91,82%	61,77%	38,23%	8,18%	65,02%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	10	73,96%	31,23%	68,77%	26,04%	70,53%	38,81%	61,19%	29,47%	65,86%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	10	72,94%	31,83%	68,17%	27,06%	72,51%	38,89%	61,11%	27,49%	66,81%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	20	93,19%	48,57%	51,43%	6,81%	86,84%	61,50%	38,50%	13,16%	62,67%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	20	92,47%	48,42%	51,58%	7,53%	86,79%	60,75%	39,25%	13,21%	63,02%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	15	93,32%	44,89%	55,11%	6,68%	85,54%	60,94%	39,06%	14,46%	62,30%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	15	92,13%	53,90%	46,10%	7,87%	89,61%	62,46%	37,54%	10,39%	63,58%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	20	75,86%	31,95%	68,05%	24,14%	69,82%	39,47%	60,53%	30,18%	65,17%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	20	74,98%	31,94%	68,06%	25,02%	71,82%	38,48%	61,53%	28,18%	66,67%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
1	35	92,85%	56,40%	43,60%	7,15%	85,17%	54,00%	46,00%	14,83%	65,59%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
1	35	92,59%	56,96%	43,04%	7,41%	87,41%	56,33%	43,67%	12,59%	65,54%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	15	93,65%	44,78%	55,22%	6,35%	83,46%	54,64%	45,36%	16,54%	64,41%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	15	91,70%	54,19%	45,81%	8,30%	85,17%	54,00%	46,00%	14,83%	65,59%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	5	75,56%	29,99%	70,01%	24,44%	72,24%	37,76%	62,24%	27,76%	67,24%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	5	76,03%	15,18%	84,82%	23,97%	72,90%	31,00%	69,00%	27,10%	70,95%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	npm	d835

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	93,02%	44,73%	55,27%	6,98%	85,34%	61,75%	38,25%	14,66%	61,82%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	92,14%	45,17%	54,83%	7,86%	84,55%	60,42%	39,58%	15,45%	62,06%

Stratified K-Fold Cross Validation with Permutation---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	94,35%	48,75%	51,25%	5,65%	89,07%	61,36%	38,64%	10,93%	63,86%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
2	35	92,14%	45,17%	54,83%	7,86%	84,55%	60,42%	39,58%	15,45%	62,06%

Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	74,99%	28,57%	71,43%	25,01%	70,67%	38,70%	61,30%	29,33%	66,01%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Gamma	Sigma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
0,1	30	74,54%	28,38%	71,62%	25,46%	69,32%	35,21%	64,79%	30,68%	67,05%

The Long Term Analysis was experimentally proved to function better than the Short Term Analysis to the separation of the outcome responses. From a first glance, one can assumes that the predicted responses at the Long Term Analysis were more clearly separable that those from the Short Term, because the methods and the input indicators in both analysis were absolutely the same. The cross validation with reproduction improved the overall accuracy and was the best validation technique when was applied to the LS-SVMs.

5.1.3.2 Long Term Analysis results with SVMs

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	96,05%	22,12%	77,88%	3,95%	74,94%	59,15%	40,85%	25,06%	57,90%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	95,37%	31,68%	68,32%	4,63%	78,05%	60,79%	39,21%	21,95%	58,63%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	95,77%	20,53%	79,47%	4,23%	72,10%	58,24%	41,76%	27,90%	56,93%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	94,90%	28,57%	71,43%	5,10%	76,30%	59,76%	40,24%	23,70%	58,27%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	7	89,92%	6,85%	93,15%	10,08%	69,90%	52,61%	47,39%	30,10%	58,64%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	7	86,22%	10,83%	89,17%	13,78%	68,96%	52,32%	47,68%	31,04%	58,32%

- ***Variables:***

Sex	Wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	97,99%	12,20%	87,80%	2,01%	73,44%	58,45%	41,55%	26,56%	57,49%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	97,12%	19,54%	80,46%	2,88%	76,03%	58,89%	41,11%	23,97%	58,57%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	10	97,94%	11,04%	88,96%	2,06%	71,27%	58,65%	41,35%	28,73%	56,31%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	10	96,90%	17,20%	82,80%	3,10%	74,10%	57,39%	42,61%	25,90%	58,36%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	94,65%	3,01%	96,99%	5,35%	70,57%	56,94%	43,06%	29,43%	56,82%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	91,54%	5,73%	94,27%	8,46%	70,18%	54,83%	45,17%	29,82%	57,67%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	25	99,48%	1,73%	98,27%	0,52%	69,53%	58,76%	41,24%	30,47%	55,38%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	25	98,96%	5,00%	95,00%	1,04%	68,85%	64,75%	35,25%	31,15%	52,05%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	2	100,00%	0,00%	100,00%	0,00%	68,62%	58,33%	41,67%	31,38%	55,15%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
6	2	100,00%	0,00%	100,00%	0,00%	67,55%	60,05%	39,95%	32,45%	53,75%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	99,07%	0,30%	99,70%	0,93%	69,54%	58,29%	41,71%	30,46%	55,63%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	97,43%	1,02%	98,98%	2,57%	69,00%	62,80%	37,20%	31,00%	53,10%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835

Stratified K-Fold Cross Validation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	99,70%	1,36%	98,64%	0,30%	67,22%	57,57%	42,43%	32,78%	54,82%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	99,11%	3,65%	96,35%	0,89%	65,34%	59,00%	41,00%	34,66%	53,17%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	1	100,00%	0,00%	100,00%	0,00%	61,75%	58,25%	41,75%	38,25%	51,75%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
5	1	100,00%	0,00%	100,00%	0,00%	64,14%	56,00%	44,00%	35,86%	54,07%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines										
TRAINING PHASE										
Parameters		Training Set---Self Test				Evaluation Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	99,35%	0,29%	99,71%	0,65%	67,30%	57,39%	42,61%	32,70%	54,95%
TESTING PHASE										
Parameters		Training Set---Self Test				Independent Test Set				
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	97,81%	0,73%	99,27%	2,19%	66,36%	56,47%	43,53%	33,64%	54,95%

➢ ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	npm	d835

Stratified K-Fold Cross Validation---Support Vector Machines											
TRAINING PHASE											
Parameters		Training Set---Self Test				Evaluation Set					
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	100,00%	0,00%	100,00%	0,00%	61,75%	58,25%	41,75%	38,25%	51,75%
TESTING PHASE											
Parameters		Training Set---Self Test				Independent Test Set					
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	1	100,00%	0,00%	100,00%	0,00%	60,00%	58,33%	41,67%	40,00%	50,83%

Stratified K-Fold Cross Validation with Permutation---Support Vector Machines											
TRAINING PHASE											
Parameters		Training Set---Self Test				Evaluation Set					
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	100,00%	0,00%	100,00%	0,00%	60,00%	58,33%	41,67%	40,00%	50,83%
TESTING PHASE											
Parameters		Training Set---Self Test				Independent Test Set					
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	15	100,00%	0,00%	100,00%	0,00%	60,00%	58,25%	41,75%	40,00%	50,88%

Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines											
TRAINING PHASE											
Parameters		Training Set---Self Test				Evaluation Set					
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	25	100,00%	0,00%	100,00%	0,00%	61,67%	57,69%	42,31%	38,33%	52,02%
TESTING PHASE											
Parameters		Training Set---Self Test				Independent Test Set					
Degree	Gamma	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
3	25	100,00%	0,00%	100,00%	0,00%	60,00%	58,33%	41,67%	40,00%	50,83%

The Support Vector Machines didn't succeed to provide a good classification performance. Despite the fact that the SVMs functioned slightly better during the Long Term than in Short term Analysis, the only improvement that was achieved was a balance to the misclassification between the uneven class sizes by increasing the specificity degree when the cross validation with permutation was operated. Conclusively, we believe that the SVMs at this study had difficulty in predicting accurately the outcome responses.

5.1.3.3 Long Term Analysis results with HS-SVMs

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol

Stratified K-Fold Cross Validation---Kernel Linear---Hidden Space Kernel Linear										
TRAINING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Evaluation Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	96,88%	65,26%	34,74%	3,12%	87,79%	71,52%	28,48%	12,21%	58,13%
TESTING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Independent Test Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	95,52%	69,58%	30,42%	4,49%	94,51%	73,68%	26,32%	5,49%	60,41%

- ***Variables:***

Sex	Wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on

Stratified K-Fold Cross Validation---Kernel Linear---Hidden Space Kernel Linear										
TRAINING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Evaluation Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	95,62%	69,44%	30,56%	4,38%	87,14%	70,03%	29,97%	12,86%	58,56%
TESTING PHASE										
Kernel Parameters	HS Kernel Parameters	Training Set---Self Test				Independent Test Set				
-	-	Sensitivity	FPR	Specificity	FNR	Sensitivity	FPR	Specificity	FNR	AUROC
-	-	95,00%	71,84%	28,16%	5,00%	95,00%	73,33%	26,67%	5,00%	60,83%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd

Stratified K-Fold Cross Validation---Kernel Linear---Hidden Space Kernel Linear										
TRAINING PHASE										
Kernel Parameters	**HS Kernel Parameters**	**Training Set---Self Test**				**Evaluation Set**				
-	**-**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**AUROC**
-	-	96,11%	67,23%	32,77%	3,89%	87,57%	72,70%	27,30%	12,43%	57,44%
TESTING PHASE										
Kernel Parameters	**HS Kernel Parameters**	**Training Set---Self Test**				**Independent Test Set**				
-	**-**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**AUROC**
-	-	94,87%	68,09%	31,91%	5,13%	93,97%	74,75%	25,25%	6,03%	59,61%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835

Stratified K-Fold Cross Validation---Kernel Linear---Hidden Space Kernel Linear										
TRAINING PHASE										
Kernel Parameters	**HS Kernel Parameters**	**Training Set---Self Test**				**Evaluation Set**				
-	**-**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**AUROC**
-	-	95,30%	59,04%	40,96%	4,70%	91,63%	67,30%	32,70%	8,37%	62,16%
TESTING PHASE										
Kernel Parameters	**HS Kernel Parameters**	**Training Set---Self Test**				**Independent Test Set**				
-	**-**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**AUROC**
-	-	93,62%	67,57%	32,43%	6,38%	91,55%	68,00%	32,00%	8,45%	61,78%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	npm	d835

Stratified K-Fold Cross Validation---Kernel Linear---Hidden Space Kernel Linear										
TRAINING PHASE										
Kernel Parameters	**HS Kernel Parameters**	**Training Set---Self Test**				**Evaluation Set**				
-	**-**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**AUROC**
-	-	93,85%	67,01%	32,99%	6,15%	93,84%	79,06%	20,94%	16,16%	57,39%
TESTING PHASE										
Kernel Parameters	**HS Kernel Parameters**	**Training Set---Self Test**				**Independent Test Set**				
-	**-**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**Sensitivity**	**FPR**	**Specificity**	**FNR**	**AUROC**
-	-	95,76%	73,19%	26,81%	4,24%	92,95%	79,58%	20,42%	7,05%	56,69%

The Hidden Space SVMs seemed to function fairly well this time. In contrast to its performance at the Short Term Analysis, the linear hidden kernel attained to separate the two classes but keeping at the same time in a low level the Specificity degree of the classifier.

5.1.3.4 Long Term Analysis results with Wrapper Method

At this part, the group with the most significant features that provided the best classification accuracy was reduced to seven. This time, the value of the AUROC was quite better, and the total number of the Support Vectors was increased. Below there is a representation of the performance of the Wrapper method after the implementation of the 20 iterations of the method. The group of features that was suggested by this method consisted of the below indicators:

PLTS_On	wbc_dia	bl_bm_dia	exm_on	PS_dia	citomol	d835

# of Features	Success Rate	AUROC	Self Accuracy	Sensitivity	Specificity	SVs (-1)	SVs (1)
10	59,55%	55,07%	91,60%	69,14%	41,00%	55	68
9	63,75%	58,74%	88,16%	74,48%	43,00%	56	68
8	66,02%	59,82%	85,02%	79,31%	40,33%	59	68
7	67,84%	61,12%	82,19%	82,24%	40,00%	61	67
6	68,75%	61,09%	79,05%	85,17%	37,00%	63	67
5	67,95%	59,92%	76,37%	85,17%	34,67%	64	72
4	67,16%	57,87%	73,14%	87,07%	28,67%	68	74
3	66,82%	56,56%	70,49%	88,79%	24,33%	70	79
2	66,36%	54,85%	68,44%	91,03%	18,67%	72	81
1	65,11%	50,52%	66,65%	96,38%	4,67%	73	88

The following table provides the participation frequency of every sample to the feature selection approach at each step, after the completion of the 20 iterations.

20 Iterations		Number of Features									
		1	2	3	4	5	6	7	8	9	10
Frequency of Every Feature	Sex	0%	0%	0%	0%	5%	5%	15%	40%	60%	100%
	wbc_dia	5%	15%	20%	45%	70%	90%	90%	100%	100%	100%
	PS_dia	10%	20%	25%	30%	45%	70%	85%	90%	100%	100%
	bl_bm_dia	10%	25%	45%	70%	80%	85%	90%	95%	100%	100%
	Hb_On	0%	10%	10%	20%	20%	25%	50%	65%	90%	100%
	PLTS_On	65%	85%	95%	95%	95%	95%	100%	100%	100%	100%
	Citomol	0%	5%	40%	55%	60%	70%	80%	90%	90%	100%
	exm_on	10%	40%	55%	65%	80%	85%	90%	95%	100%	100%
	Itd	0%	0%	0%	5%	10%	35%	45%	70%	80%	100%
	d835	0%	0%	10%	15%	35%	40%	55%	55%	80%	100%

5.1.3.5 Long Term Analysis results with Classifier Fusion

Following the same formulation as in chapter 5.1.2.5, the Level 2 classification results are given below. Note that this time some of the classifier ensemble techniques provided equivalent performance with the performance of the LS-SVM classifier.

Classifier Fusion based on the best AUROC	Level 2 Classifier	Combination of L1 Cls	Sensitivity	FPR	Specificity	FNR	AUROC	Diversity
	Majority Vote	NBC--HSSVM--LSSVM	73,52%	31,77%	68,23%	26,48%	70,88%	0,329
	Minimum	RBNC--HSSVM--LSSVM	73,28%	33,00%	67,00%	26,72%	70,14%	0,617
	Maximum	RBNC--HSSVM--LSSVM	73,28%	33,00%	67,00%	26,72%	70,14%	0,617
	Average	LDC--HSSVM--LSSVM	72,93%	32,67%	67,33%	27,07%	70,13%	0,222
	Product Rule	LDC--HSSVM--LSSVM	72,93%	32,67%	67,33%	27,07%	70,13%	0,222
	Decision Templates	KNNC--LDC--QDC RNNC--Fisher—HSSVM LSSVM	74,14%	33,33%	66,67%	25,86%	70,40%	0,391
	Fisher	LDC--HSSVM	79,14%	43,67%	56,33%	20,86%	67,74%	0,229
	KNN	LDC--QDC--Fisher HSSVM--LSSVM	82,07%	47,33%	52,67%	17,93%	67,37%	0,316
	LDC	KNNC--HSSVM	79,31%	44,32%	55,68%	20,69%	67,50%	0,531
	QDC	RNNC--LSSVM	69,14%	46,00%	54,00%	30,86%	61,57%	0,514
	RBNC	LDC--QDC--Fisher	80,17%	51,00%	49,00%	19,83%	64,59%	0,356
	RNNC	LDC--QDC--Fisher--HSSVM	82,93%	48,67%	51,33%	17,07%	67,13%	0,332

Classifier Fusion based on the best averaged Diversity	Level 2 Classifier	Combination of L1 Cls	Sensitivity	FPR	Specificity	FNR	AUROC	Diversity
	Majority Vote	RBNC--LSSVM	83,49%	62,33%	37,67%	16,51%	60,58%	0,847
	Minimum	RBNC--LSSVM	75,35%	36,33%	63,67%	24,65%	69,51%	0,847
	Maximum	RBNC--LSSVM	75,35%	36,33%	63,67%	24,65%	69,51%	0,847
	Average	RBNC--LSSVM	75,35%	36,33%	63,67%	24,65%	69,51%	0,847
	Product Rule	RBNC--LSSVM	75,35%	36,33%	63,67%	24,65%	69,51%	0,847
	Decision Templates	RBNC--LSSVM	69,31%	29,33%	70,67%	30,69%	69,99%	0,847
	Fisher	RBNC--LSSVM	82,41%	53,67%	46,33%	17,59%	64,37%	0,847
	KNN	RBNC--LSSVM	81,55%	55,00%	45,00%	18,45%	63,28%	0,847
	LDC	RBNC--LSSVM	79,66%	50,67%	49,33%	20,34%	64,50%	0,847
	QDC	RBNC--LSSVM	76,72%	70,67%	29,33%	23,28%	53,03%	0,847
	RBNC	RBNC--LSSVM	81,55%	53,67%	46,33%	18,45%	63,94%	0,847
	RNNC	RBNC--LSSVM	82,24%	56,67%	43,33%	17,76%	62,79%	0,847

Classifier Fusion with all the available L1 classifiers	Level 2 Classifier	Combination of L1 Cls	Sensitivity	FPR	Specificity	FNR	AUROC	Diversity
	Majority Vote	all L1 classifiers	81,21%	49,00%	51,00%	18,79%	66,11%	0,448
	Minimum	all L1 classifiers	78,79%	43,00%	57,00%	21,21%	67,90%	0,448
	Maximum	all L1 classifiers	77,93%	42,67%	57,33%	22,07%	67,63%	0,448
	Average	all L1 classifiers	78,62%	42,67%	57,33%	21,38%	67,98%	0,448
	Product Rule	all L1 classifiers	77,93%	42,00%	58,00%	22,07%	67,97%	0,448
	Decision Templates	all L1 classifiers	73,45%	34,00%	66,00%	26,55%	69,73%	0,448
	Fisher	all L1 classifiers	80,69%	55,33%	44,67%	19,31%	62,68%	0,448
	KNN	all L1 classifiers	80,00%	53,00%	47,00%	20,00%	63,50%	0,448
	LDC	all L1 classifiers	62,07%	53,00%	47,00%	37,93%	54,54%	0,448
	QDC	all L1 classifiers	38,79%	35,33%	64,67%	61,21%	51,73%	0,448
	RBNC	all L1 classifiers	99,48%	100%	0,00%	0,52%	49,74%	0,448
	RNNC	all L1 classifiers	78,10%	52,00%	48,00%	21,90%	63,05%	0,448

5.1.3.6 Discussion to the Long Term Analysis Results

Before starting with the representation of the classification results, we should mention that the Long Term Analysis was an easier classification problem with a much better averaged classification performance than the performance provided by classifiers form the Short Term Analysis. We believe that this result was caused by the fact that the two classes had now a more clear-cut meaning (Complete Remission versus Resistance and Death from disease), and therefore some sharply discriminated characteristics separating these two outcomes.

Our group has also detected, for a second time, the improvement to the performance of the classifiers by our alternative cross validation techniques. Furthermore, the highly used cross validation techniques corresponded better this time to the

classification problem. Almost all the classifiers achieved to correctly classify a high percentage of the unseen samples from the dense class (high Sensitivity) and to keep the Specificity of the classifier close, and sometimes above 40%.

The Hidden Space SVMs with a linear Hidden Kernel hopefully succeeded to attain a good Specificity but kept the level of the Specificity quite low. The Support Vector Machines provided a lightly worse result and our interest was now focused on the total number of the Support Vectors that each classifier needed for the separation of the two classes. The feature selection classification accuracy, provided no significant difference and its performance was equal the same with the performance of the SVMs and the HS-SVM classifier.

Finally, when our alternative cross validation techniques implemented to the Support Vector Machines, we observed no further improvement to the area under the ROC curve. The only reaction was the more balanced values of the two measures. Particularly, the cross validation with reproduction, tended more than others to narrow the wide between the high value of the Sensitivity and the low-valued Specificity. The conclusion arising by these results was that the influence of the alternative cross validation techniques to the performance of the SVM classifier was reduced as the gap between the distribution of the two classes was narrowed.

Stratified K-Fold Cross Validation---Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
78,05%	60,79%	39,21%	21,95%	58,63%
Stratified K-Fold Cross Validation with Reproduction---Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
68,96%	52,32%	47,68%	31,04%	58,32%
Stratified K-Fold Cross Validation with Permutation---Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
76,30%	59,76%	40,24%	23,70%	58,27%
Stratified K-Fold Cross Validation---Kernel Linear---Hidden Space Kernel Linear				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
91,55%	68,00%	32,00%	8,45%	61,78%
Wrapper Technique				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
82,24%	60,00%	40,00%	17,76%	61,12%

On the other hand, the LS-SVM was also this time the most accurate classifier. Particularly, when it was trained according to the stratified K-fold cross validation with reproduction, the area under the ROC curve achieved the highest percentage that we had seen at the examination of the AML99 dataset. The reproduction of the data tended to increase the AUROC, keeping at the same time both the Sensitivity and Specificity in a high value.

Stratified K-Fold Cross Validation---Least Squares Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
87,41%	56,33%	43,67%	12,59%	65,54%
Stratified K-Fold Cross Validation with Reproduction---Least Squares Support Vector Machines				
Sensitivity	**FPR**	**Specificity**	**FNR**	**AUROC**
72,90%	31,00%	69,00%	27,10%	70,95%

To sum up, the conclusion that was made from the Long Term analysis examination was that the Level 1 LS-SVM was confirmed as the most powerful classifier when it was trained with equivalent amount of information from both classes. It is very important to report that the dataset that provided the best classification was consisted of the same variables that achieved the highest performance at the Short Term Analysis. Such dataset consisted of indicators:

Sex	**wbc_dia**	**PS_dia**	**bl_bm_dia**	**Hb_On**	**PLTS_On**	**citomol**	**exm_on**	**itd**	**d835**

The classifier fusion approach provided no further improvement to the classification performance. This was an issue that worried us a lot because such technique generally provides a better separation between the classes than the individual classifiers. We believe that these results derived from the low performance of almost all the Level 1 classifiers, except for the LS-SVMs, and generally the high correlation among them.

Another issue for discussion was the number of the Support Vectors that required for the SVMs and HS-SVMs procedures. Hopefully, the Hidden Space SVMs classified fairly well this time, and a comparison between these two different formulations is given below. Despite the fact that the best classification accuracy was achieved by a certain dataset, the following tables give a thorough presentation of all the applicable Kernel and

Hidden Kernel functions when applied to the datasets, given by Table 25. Our experiments were also based on several Rbf Kernels with different parameters, and for that reason the parameter of the Rbf Kernel was also given. Note also that the polynomial Hidden Kernel totally failed to separate the classes and for that reason is not presented below.

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol

Number of Samples : 359 --- Samples from (-1) : 114 --- Samples from (1) : 245					
Kernel	**HS Kernel**	**# of SVs (-1)**	**# of SVs (1)**	**# of Total SVs**	**AUROC**
-	Linear	89	91	180	60,41%
-	Rbf with gam=2	85	95	180	58,51%
-	Rbf with gam=5	88	94	182	58,39%
Linear	-	91	93	184	60,35%
-	Rbf with gam=1	85	100	185	58,40%
-	Rbf with gam=10	92	97	189	58,01%
Rbf with gam=5	-	92	101	193	59,25%
Rbf with gam=10	-	94	100	194	59,23%
Rbf with gam=2	-	91	108	199	60,24%
Rbf with gam=1	-	91	118	209	58,69%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on

Number of Samples : 335 --- Samples from (-1) : 105 --- Samples from (1) : 230					
Kernel	**HS Kernel**	**# of SVs (-1)**	**# of SVs (1)**	**# of Total SVs**	**AUROC**
-	Linear	82	83	165	60,83%
Linear	-	84	86	170	60,83%
-	Rbf with gam=2	80	91	171	57,05%
-	Rbf with gam=5	83	90	173	58,87%
-	Rbf with gam=1	79	96	175	56,35%
-	Rbf with gam=10	85	90	175	57,12%
Rbf with gam=10	-	87	93	180	58,96%
Rbf with gam=5	-	86	95	181	60,18%
Rbf with gam=2	-	85	105	190	58,56%
Rbf with gam=1	-	85	115	200	55,84%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd

Number of Samples : 289 --- Samples from (-1) : 96 --- Samples from (1) : 193					
Kernel	**HS Kernel**	**# of SVs (-1)**	**# of SVs (1)**	**# of Total SVs**	**AUROC**
-	Linear	69	71	140	59,61%
Linear	-	71	73	144	59,61%
-	Rbf with gam=5	71	77	148	55,13%
-	Rbf with gam=2	70	79	149	54,27%
-	Rbf with gam=10	73	77	150	56,66%
-	Rbf with gam=1	71	86	157	53,59%
Rbf with gam=10	-	75	82	157	56,96%
Rbf with gam=5	-	75	83	158	55,95%
Rbf with gam=2	-	74	91	165	55,38%
Rbf with gam=1	-	75	102	177	55,70%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	d835

Number of Samples : 259 --- Samples from (-1) : 89 --- Samples from (1) : 170					
Kernel	**HS Kernel**	**# of SVs (-1)**	**# of SVs (1)**	**# of Total SVs**	**AUROC**
-	Linear	67	69	136	61,78%
Linear	-	69	71	140	61,50%
-	Rbf with gam=5	69	75	144	55,45%
-	Rbf with gam=10	71	75	146	54,25%
-	Rbf with gam=2	68	80	148	52,98%
Rbf with gam=10	-	73	80	153	57,54%
Rbf with gam=5	-	72	82	154	56,36%
-	Rbf with gam=1	68	87	155	52,27%
Rbf with gam=2	-	71	90	161	54,47%
Rbf with gam=1	-	71	100	171	52,83%

- ***Variables:***

Sex	wbc_dia	PS_dia	bl_bm_dia	Hb_On	PLTS_On	citomol	exm_on	itd	npm	d835

Number of Samples : 197 --- Samples from (-1) : 70 --- Samples from (1) : 127					
Kernel	**HS Kernel**	**# of SVs (-1)**	**# of SVs (1)**	**# of Total SVs**	**AUROC**
-	Linear	51	54	105	56,69%
Linear	-	54	56	110	56,34%
-	Rbf with gam=5	54	61	115	53,20%
-	Rbf with gam=10	56	60	116	53,18%
-	Rbf with gam=2	53	67	120	51,48%
Rbf with gam=10	-	58	66	124	54,22%
Rbf with gam=5	-	56	70	126	53,66%
Rbf with gam=2	-	56	82	138	52,73%
Rbf with gam=1	-	57	91	148	50,87%
-	Rbf with gam=1	85	100	185	51,89%

The conclusion obtained from the above results was that in general terms the Hidden Space SVMs required less Support Vectors than the simple SVMs to classify, with almost the same performance, the two classes. The linear Hidden Space Kernel provided the best classification accuracy but without a concerning reduction to the Support Vectors when compared to the simple linear Kernel application. On the other hand, the comparison between the Hidden and the simple Rbf Kernel was very representative and experimentally confirmed Zhang's observation.

5.2 Survival Analysis Results

Below, we represent the survival analysis results from the Short Term and Long Term Analysis, with respect to the Kaplan-Meier survival analysis **[71]**. A plot of the Kaplan-Meier estimate of the survival function is a series of horizontal steps of declining magnitude which, when a large enough sample is taken, approaches the true survival function for that population. The value of the survival function between successive distinct sampled observations ("clicks") is assumed to be constant. An important advantage of the Kaplan-Meier curve is that the method can take into account "censored" data, losses from the sample before the final outcome is observed (for instance, if a patient withdraws from a study). On the plot, small vertical tick-marks indicate losses, where patient data has been censored.

In order to combine the pattern recognition with the statistical analysis, we focused on the application of the survival analysis to the best classification results. The classifier with the most accurate results in both cases was the LS-SVM with Rbf Kernel, trained with the stratified K-fold cross validation with reproduction.

As noticed above, the assessment of the classification accuracy was defined by the iterated testing procedure, in which 20 randomly bootstrapped testing and training sets determined the overall performance. For that reason, the survival analysis results had to deal with the 20 different testing results from the classification. The solution that was recommended by our group was to present the survival analysis from the iteration that provided the best classification accuracy, the lowest area under the ROC curve, and finally to combine all the results from the 20 iterations into a unique dataset and proceed with the survival analysis of it. We adopted this idea in order to represent in a clearly way the survival analysis that gathered from all the levels of performance from a specified classifier, from the lowest to the highest classification accuracy of it.

5.2.1 Short Term Survival Analysis Results

- **LS-SVM classification results with maximum accuracy**

Case Processing Summary

Predicted	Total N	N of Events	Censored	
			N	Percent
-1	12	5	7	58,3%
1	32	0	32	100,0%
Overall	44	5	39	88,6%

No statistics about the Confidence Interval and the STD Error were computed because all cases were censored. Moreover, one can easily observe that the cumulative survival curve correspond to class -1 does not start from the starting point (0, 1). This issue is analytically explained in the following chapter 5.2.3.

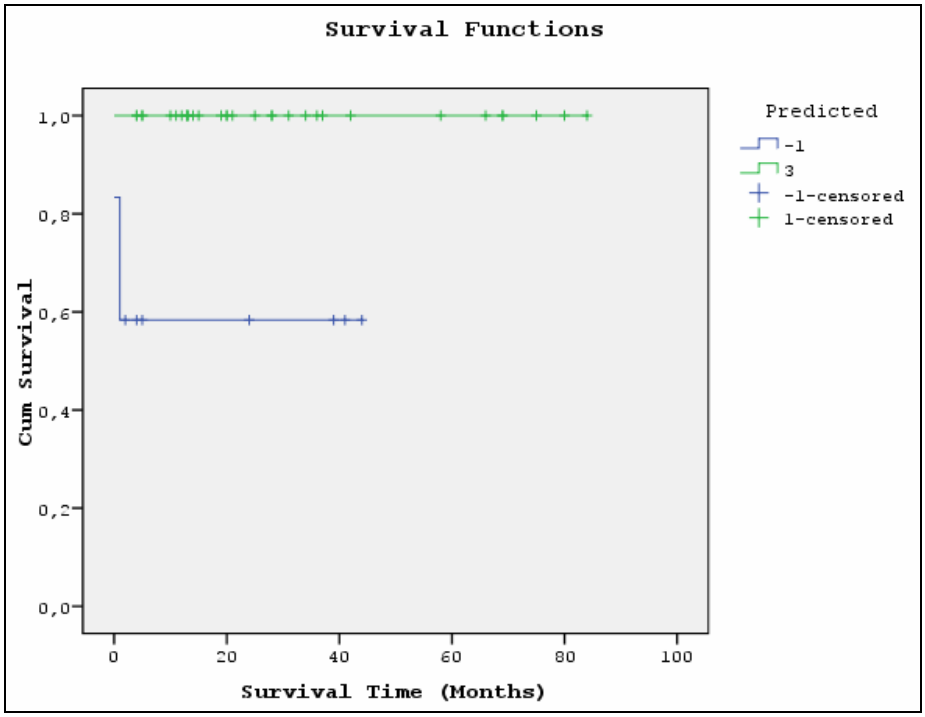

Figure 26 Short Term Survival Analysis by maximum classification accuracy

- **LS-SVM classification results with minimum accuracy**

Case Processing Summary

Predicted	Total N	N of Events	Censored	
			N	Percent
-1	11	1	10	90,9%
1	33	4	29	87,9%
Overall	44	5	39	88,6%

Predicted	Mean (Estimation is limited to the largest survival time if it is censored)			
	Estimate	Std. Error	95% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
-1	31,82	3,03	25,87	37,76
1	74,82	4,77	65,46	84,17
Overall	75,43	4,03	67,54	83,33

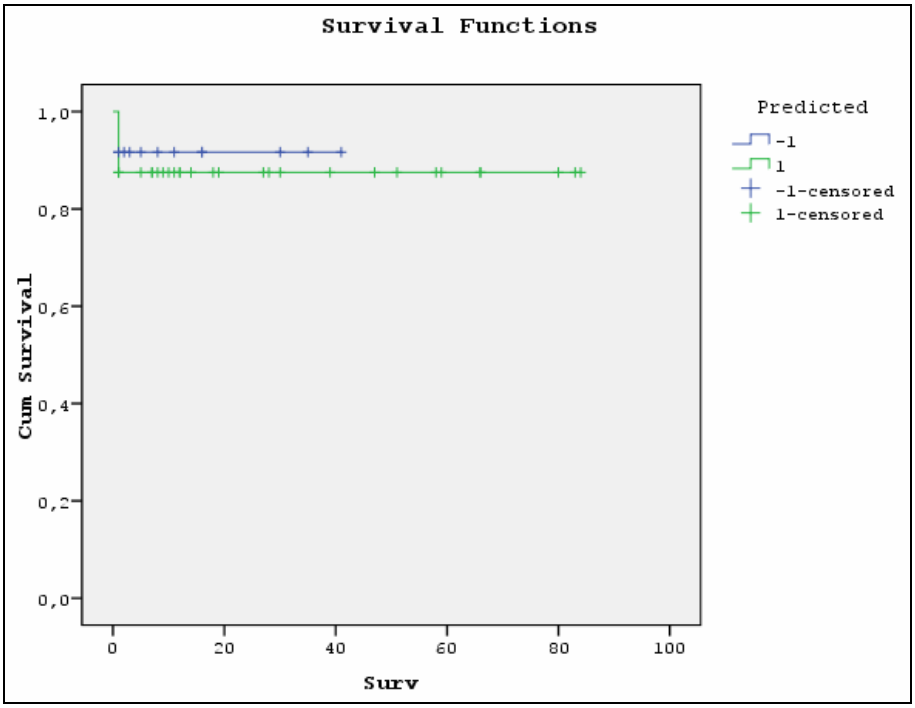

Figure 27 Short Term Survival Analysis by minimum classification accuracy

- **Combination of all the LS-SVM classification results**

Case Processing Summary

Predicted	Total N	N of Events	Censored	
			N	Percent
-1	260	60	200	76,9%
1	620	40	580	93,5%
Overall	880	100	780	88,6%

Predicted	Mean (Estimation is limited to the largest survival time if it is censored)			
	Estimate	Std. Error	95% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
-1	55,54	1,86	51,89	59,19
1	79,58	0,83	77,96	81,21
Overall	75,43	0,90	73,67	77,20

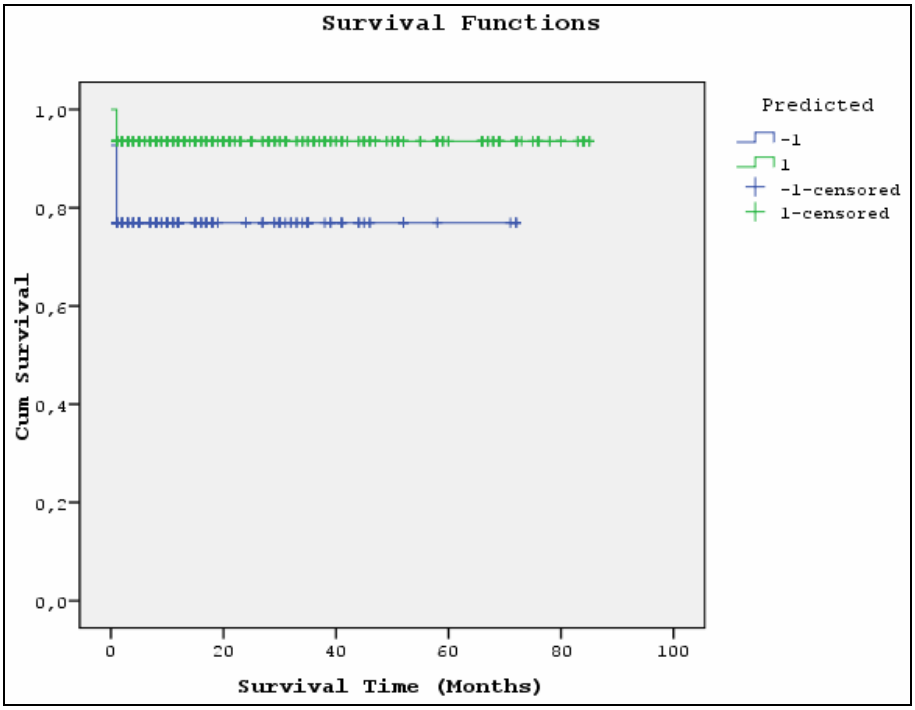

Figure 28 Short Term Survival Analysis by all the classification accuracies

5.2.2 Long Term Survival Analysis Results

- **LS-SVM classification results with maximum accuracy**

Case Processing Summary

predict	Total N	N of Events	Censored	
			N	Percent
-1	15	11	4	26,7%
1	29	4	25	86,2%
Overall	44	15	29	65,9%

Predicted	Mean (Estimation is limited to the largest survival time if it is censored)			
	Estimate	Std. Error	95% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
-1	17,47	4,36	8,93	26,01
1	65,00	4,35	56,46	73,54
Overall	48,97	5,01	39,15	58,78

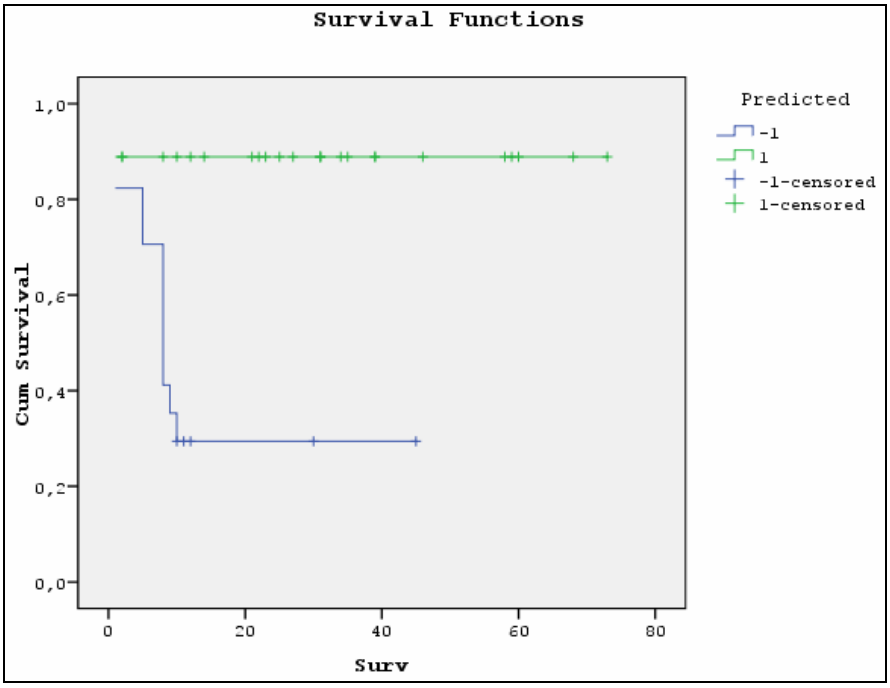

Figure 29 Long Term Survival Analysis by maximum classification accuracy

- **LS-SVM classification results with minimum accuracy**

Case Processing Summary

Predicted	Total N	N of Events	Censored	
			N	Percent
-1	23	10	13	56,5%
1	21	5	16	76,2%
Overall	44	15	29	65,9%

Predicted	Mean (Estimation is limited to the largest survival time if it is censored)			
	Estimate	Std. Error	95% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
-1	44,74	8,31	28,46	61,03
1	64,38	7,66	49,37	79,39
Overall	55,77	5,91	44,19	67,35

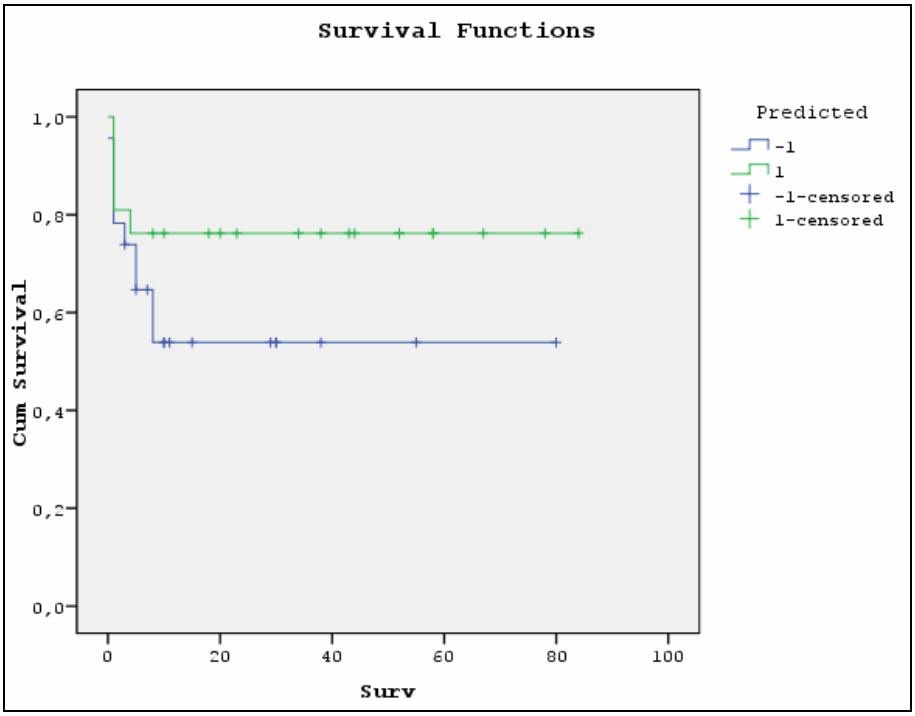

Figure 30 Long Term Survival Analysis by minimum classification accuracy

- **Combination of all the LS-SVM classification results**

Case Processing Summary

pred	Total N	N of Events	Censored	
			N	Percent
-1	369	206	163	44,2%
1	511	94	417	81,6%
Overall	880	300	580	65,9%

Predicted	Mean (Estimation is limited to the largest survival time if it is censored)			
	Estimate	Std. Error	95% Confidence Interval	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
-1	36,24	2,00	32,31	40,17
1	68,79	1,52	65,82	71,76
Overall	55,64	1,38	52,93	58,34

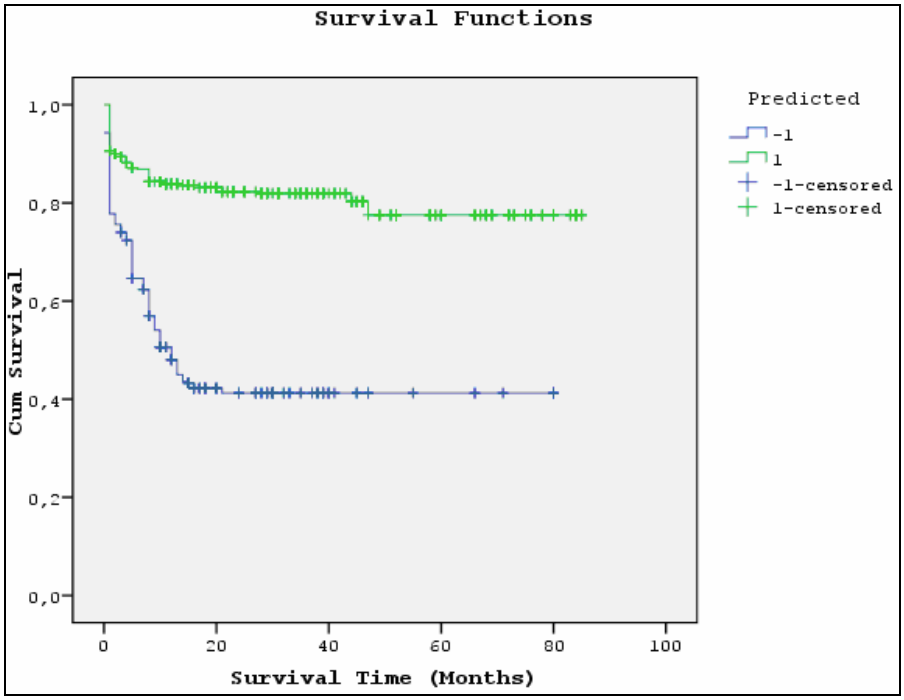

Figure 31 Long Term Survival Analysis by all the classification accuracies

5.2.3 Discussion to the Survival Analysis results

From the above survival analysis figures we observe that a good classification performance generally leads to a good separation between the survival curves of the two outcome responses. Specifically, we experimentally proved that the LS-SVM was the best classification procedure in both Short and Long Term analysis. As a result, the survival analysis figures that provided from this classifier, especially after the Long Term Analysis, presented a clear difference between the cumulative survival curves of the two responses.

Another observation that needs discussion is the fact that in some cases the survival curves do not start from the top of the figure. Generally, the vertical (or Y) axis gives the proportion of people surviving. The value is a fraction which runs from 1 at the top to the zero at the bottom, representing 100% survival to zero percent survival at the bottom. For that reason one could assume that all figures, no matter what is the outcome response of the patients, must start from point (0, 1). Below we give an analytical explanation of this issue.

The survival function $S(t)$ is defined as the probability that a person survives at a specified time t. As a result, at time $t = 0$ where no events occur the value of $S(t)$ is equal to 1 and the survival curve starts from point (0, 1). In our dataset, the variable that indicates the survival time in months is normalized and all the patients with survival time less or equal to 0.5 are normalized to zero. According to this, when a patient failures to survive at time $t = 0.4$, practically at time $t = 0$, then an event occurs and the value of the survival function $S(t)$ is less than one.

6 General Discussion and Further Work

Summing up with some general discussions about the AML99 analysis, we should confess that the current dataset was a very hard discriminated dataset. As far as we know, this was the first time that this dataset was examined from the pattern recognition field of study and we are very happy about the demanding research that was partially accomplished by our team. A lot of attention has been paid, by our coordinators so far, to the statistical analysis of the AML99 data. This study aimed to complement the hard work that has been done for the design and the implementation of a good and reliable prognostic system.

The knowledge that could be offered to whom it may concerned about the pattern recognition application to the AML99 dataset, is that several modern and highly-used classification techniques were applied to such dataset. Moreover, several extensions and alternative methodologies were designed and implemented in every step of the classification approach. Our group was involved with the structure of the classification and suggested new different techniques, from the validation of the dataset during the classification, to the development of new methodologies for classification. We are currently looking forward to publishing some of our researches through the evaluation of them by several other biomedical datasets.

Although fairly good results were finally obtained, some open issues remained. These are outlined below with possible indications on how they might be solved. Getting started with the preparation of the data for classification, some further studies could be

applied. The tested indicators could be examined from the statistical point of view by assessing their correlation degree between them. This study should be possible provide relevant information about the significance of every indicator that could be afterwards combined and compared with the significant group of features, provided by the Wrapper technique.

What is more, some open issues remained from the Level 1 classification and precisely from the application of our alternative cross validation methods. Having in mind that generally the alternative cross validation approaches contributed to an optimized classification accuracy, one should wondered why such methods were not applied to the Hidden Space SVMs and the Wrapper technique.

As for the case of the Hidden Space SVM, our group was focused more on the evaluation of the Zhang's observation about the reduction of the Support Vectors when such techniques were adopted, rather than to attempt to improve the performance of this classifier. We tried to simulate the classification procedure with almost the same way that Zhang did to his research and see the same reaction from the AML99 dataset. For that reason, the differentiation to the cross validation techniques to the Hidden Spaces SVMs, could produce no comparable results.

In case of the Wrapper technique, as noticed from the beginning, the design and implementation of this approach was provided by a PhD research. For that reason, a possible extension could be proved as a very time consuming because this study had to interfere to the software, and the structure of this method. Referring to the Wrapper method, one possible approach worth of further investigation was the experimentation with the classifier that this method adopted. An interesting study would be the application of several classifiers and precisely the LS-SVMs to the feature selection. We believe that this change would possible produce good classification results.

An additional natural extension of our method could be to verify the improvement of the alternative cross validation techniques to several other datasets. We previously noticed that the problem with the non well-distributed classes was quite common to latest researches, and a possible application of these methods to such datasets could be worthwhile. Furthermore, it will be really interesting to compare the classifier

performance from our alternative cross validation techniques with the method presented in **[68]**.

Finally, the most important study in our point of view, could definitely be the theoretically verification of the improvement of our cross validation approaches to the SVMs and the LS-SVM classifier. We observed that different reactions occurred when the reproduction and the permutation of the data contributed to the cross validation of the SVM and the LS_SVM classifier, respectively. In case of the LS-SVM, an improvement to the AUROC was also gained. Our group has paid a lot of attention to this study, and already been working on this field, and we hopefully believe that very soon we will have the chance to confirm these results.

7 References

[1] Cytomorphology of acute myeloid leukemia and myelodysplastic syndromes. Rinsho Ketsueki. Pubmed, 2006 August; 47 (8) 701-9.

[2] AML1 and the 8;21 and 3;21 translocations in acute and chronic myeloid leukemia. G Nucifora, J D Rowley, Blood 1995 July 1;86 (1): 1-14.

[3] The Importance of Diagnostic Cytogenetics on Outcome in AML: Analysis of 1,612 Patients Entered into the MRC AML 10 Trial, By David Grimwade, Blood Vol 92, No 7 (October 1), 1998: 2322-2333.

[4] Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. T R Golub , D K Slonim , P Tamayo , C Huard , M Gaasenbeek , J P Mesirov , H Coller , M L Loh , J R Downing , M A Caligiuri , C D Bloomfield , E S Lander. Science 1999 Oct 15; 286 (5439): 531-7.

[5] MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Scott A Armstrong , Jane E Staunton , Lewis B Silverman , Rob Pieters , Monique L den Boer , Mark D Minden , Stephen E Sallan , Eric S Lander, Todd R Golub , Stanley J Korsmeyer. Nat Genet. 2002 Jan; 30 (1):41-7.

[6] Support vector machine classification and validation of cancer tissue samples using microarray expression data. T S Furey , N Cristianini , N Duffy , D W Bednarski , M Schummer, D Haussler Bioinformatics. 2000 Oct; 16 (10):906-14.

[7] Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. Lars Bullinger , Konstanze Döhner , Eric Bair , Stefan Fröhling , Richard F Schlenk , Robert Tibshirani , Hartmut Döhner , Jonathan R Pollack N Engl J Med. 2004 Apr 15;350 (16):1605-16.

[8] www.**gimema**.org

[9] Ilaria Ardoino, Simona Iacobelli, Federico Ambrogi, Patrizia Boracchi and Elia Biganzoli, Biopattern's report.

[10] Keith Roberts, Martin Raff, Bruce Alberts, Peter Walter, Julian Lewis and Alexander Johnson, Molecular Biology of the Cell. 28/02/2002, ISBN 0815332181-Hardback book.

[11] Levitsky G.A. 1931. The morphology of chromosomes. Bull. Applied Bot. Genet. Plant Breed. **27**, 19-174.

[12] http://www.medicinenet.com

[13] http://www.cancerbackup.org.uk

[14] www.cancer.org

[15] http://bloodjournal.hematologylibrary.org/

[16] http://www.lrf.org.uk

[17] http://www.inctr.org/

[18] http://www.slh.wisc.edu/wps/wcm/connect/extranet/home/

[19] Clinical Importance of Cytogenetics in Acute Myeloid Leukaemia. Best Practice & Research Clinical Haematology, Volume 14, Issue 1, Pages 19-47 K. Mrózek

[20] www.genenames.org

[21] www.atlasgeneticsoncology.org

[22] Advanced age and high initial WBC influence the outcome of inv(3)(q21q26)/t(3;3) (q21;q26) positive AML. Martin Weisser, Claudia Haferlach , Torsten Haferlach, Susanne Schnittger. Leukemia and Lymphoma, Volume 48, Issue 11 November 2007 , pages 2145 – 2151.

[23] Correlation between karyotype and quantitative immunophenotype in acute myelogenous leukemia with t(8;21), Modern Pathology (2004) **17,** 1211–1216, advance online publication, 4 June 2004, Haytham Khoury, Bakul I Dalal, Stephen H Nantel.

[24] Erickson PF, Robinson M, Owens G, *et al*. The ETO portion of acute myeloid leukemia t(8;21) fusion transcript encodes a highly evolutionarily conserved, putative transcription factor. *Cancer Res* 1994;54:1782–1786.

[25] Kurzrock R, Kantarjian HM, Druker BJ, Talpaz M. "Philadelphia Chromosome-positive leukemias: from basic mechanisms to molecular therapeutics." *Ann Intern Med* 2003;138:819–30. PMID 12755554.

[26] Analysis of DEK-CAN fusion gene expression in acute myeloid leukemia patients with 6; 9 chromosome translocation. Zhongguo Shi Yan Xue Ye Xue Za Zhi. 2006 Apr ;14 (2):232-6 16638187, Pubmed.

[27] Molecular identification of CBF^{β}-MYH11 fusion transcripts in an AML M4Eo patient in the absence of inv16 or other abnormality by cytogenetic and FISH analyses – a rare occurrence. F Ravandi, S S Kadkol, J Ridgeway, A Bruno, C Dodge and V Lindgren. Leukemia (2003) 17, 1907–1910.

[28] Chromosomal Aberration of the 11q23 Locus in Acute Leukemia and Frequency of MLL Gene Translocation. M. Christina Cox, Paola Panetta, 2004 American Society for Clinical Pathology.

[29] The journal of Molecular Diagnostics. http://jmd.amjpathol.org/.

[30] Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values, Tapio Schneider. American Meteorological Society.

[31] Dimou Ioannis, Manikis Georgios, Zervakis Michalis. Classifier Fusion Approaches for Diagnostic Cancer Models. 28th IEEE EMBS, September 2006, New York

[32] I. Dimou, B. Van Calster, S. Van Huffel, D. Timmerman, M. Zervakis: Comparison of Imputation Approaches In Ovarian Tumor Diagnostic Models Based On LS-SVMs, *EMBC2007 Proceedings,* 29th IEEE EMBS Annual International Conference, August 23-26, 2007.

[33] Pai-Hsuen Chen, Chih-Jen Lin and Bernhard Scholkopf, "A Tutorial on ν-Support Vector Machines". Applied Stochastic Models in Business and Industry, Vol. 21, No. 2. (2005), pp. 111-136.

[34] Cortes C., Vapnik V. (1995), "Support Vector Networks" Machine Learning, 20, 273-297.

[35] Suykens J.A.K., Vandewalle J. (1999) "Least squares support vector machine classifiers", Neural Processing Letters, 9(3), 293-300.

[36] Li Zhang, Weida Zhou and Licheng Jiao, "Hidden Space Support Vector Machines", *IEEE Transactions on Neural Networks*, vol. 15, no. 6, pp 1424-1434, 2004.

[37] L. I. Kuncheva, J. C. Bezdek, R. P. W. Duin, "Decision templates for multiple classifier fusion: An experimental comparison", Pattern Recognition, 34, (2), 2001.

[38] H. Altıncay, "On naive Bayesian fusion of dependent classifiers.", Pattern Recognition Letters (2005).

[39] K. Tumer, J. Ghosh, "Classifier combining: Analytical results and implications". In Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, workshop at the Thirteenth National Conference on Artificial Intelligence, Portland, OR, August 1996.

[40] L.I.Kuncheva, "A theoretical study on six classifier fusion strategies",IEEE Trans. on Pattern Analysis and Machine Intelligence,Vol.24, No.2, February 2002.

[41] D. Ruta, B.Gabrys, "an overview of classifier fusion methods", Computing and Information Systems, 7 (2000) p.1-10.

[42] K. Woods, W. Ph. Kegelmeyrer, K. Bowyer, "Combination of multiple classifiers using local accuracy estimates", IEEE transactions on patern analysis and machine intelligence, vol. 19, no. 4, april 1997.

[43] G. Fumera, F. Roli, "Linear combiners for classifier fusion: some theoretical and experimental results". IEEE Transactions on Pattern Analysis and Machine Intelligence archive. Volume 27, Issue 6 (June 2005). Pages: 942 – 956.

[44] L.I.Kuncheva, L.C.Jain, "Designing classifier fusion systems by genetic algorithms."IEEE Transactions on Evolutionary Computation*,4(4), 2000, 327-336*

[45] J. Kittler, M. Hatef, R. P. W. Duin, J. Matas, "On combining classifiers", IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 3, March 1998

[46] K. Chen, L. Wang, H. Chi, "Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification", International Journal of Pattern Recognition and Artificial Intelligence, 11(3), 1997, pp. 417-445

[47] Combining Pattern Classifiers, Methods and Algorithms, Ludmila I. Kuncheva. John Wiley & Sons, 2004.

[48] G.U. Yule. On the association of attributes in statistics. Phil. Trans., A, 194, (1900), 257-319

[49] T.K. Ho. The random space method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), (1998), 832-844.

[50] D.B. Skalak. The sources of increased accuracy for two proposed boosting algorithms. The Journal of Machine Learning Research. Volume 5, (December 2004). Pages: 421 – 451.

[51] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification. Image and Vision Computing, Volume 19, Number 9, 1 August 2001 , pp. 699-707(9).

[52] Michalis Petrakos, Jon Atli Benediktsson, Senior Member IEEE, and Ioannis Kanellopoulos, Member IEEE. The effect of Classifier Agreement on the Accuracy of the Combined Classifier in Decision Level Fusion. IEEE Transactions on Geoscience and Remote Sensing, Vol. 39, NO. 11, November 2001

[53] R. Kohavi and D.H. Wolpert. Bias plus variance decomposition for zero-one loss functions. in L. Saitta, (Ed.), Machine Learning: Proc. 13th International Conference, Morgan Kaufmann, 1996, pp. 275-283.

[54] L. Xu, A. Krzyzak, and C.Y. Suen. Methods of combining multiple classifiers and their application to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics, 22:418{435, 1992.

[55] R.F. Bordley. A multiplicative formula for aggregating probability assessments. Management Science, 28:1137-1148, 1982.

[56] L. I. Kuncheva. “Fuzzy” vs. “non-fuzzy” in combining classifiers designed by boosting. IEEE Transactions on Fuzzy Systems 11:729–741, 2003.

[57] L. I. Kuncheva. Using measures of similarity and inclusion for multiple classifier fusion by decision templates. Fuzzy Sets and Systems, 122(3):401–407, 2001.

[58] G. Rogova. Combining the results of several neural network classifiers. Neural Networks, 7:777{781,1994.

[59] R.O. Duda and P.E. Hart. Pattern Classification and Scene Analysis. John Wiley & Sons, NY, 1973.

[60] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. Artificial Intelligence, 97:245–271, 1997.

[61] The MathWorks- Matlab and Simulink for Technical Computing.

[62] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002 (ISBN 981-238-151-1).

[63] The OSU SVM Support Vector Machine Toolbox for MATLAB is Copyright (C) 2000-2003 by Chih-Chung Chang and Chih-Jen Lin under the BSD License.

[64] www.spss.com

[65] http://biopattern.uninova.pt/

[66] Improving filter methods through wrapper approaches using kernels-a breast cancer case study, M. Blazantonakis, M Zervakis. Not published yet.

[67] http://www.comp.rgu.ac.uk/staff/chb/teach.html

[68] Weighted support vector machine for classification, Shu-Xin Du, Sheng-Tan Chen. Proceedings of the forth international conference on Machine Learning and Cybernetics, Guangzhou, 18-21 August 2005.

[69] http://prtools.org/

[70] S. Armstrong, J. Staunton, L. Silverman, R. Pieters, et al. “MLL translocations specify a distinct gene expression profile that distinguishes a unique leukaemia”, *nature genetics*, **30**, pp. 41-47 (2002).

[71] Kaplan, E, Meier, P. Nonparametric estimation from incomplete observations. Am J Stat Assoc. 1958;53:457-481.